ANÁLISIS DE LA CYBER-RIVALIDAD Y HOSTILIDAD EN REDES SOCIALES. CASO DE ESTUDIO: TWITTER.

19 de Junio de 2017

AUTORA

PATRICIA RODRÍGUEZ MONTERO

TUTOR

JOSE MARÍA ÁLVAREZ RODRÍGUEZ

Grado en Ingeniería Informática (Plan 2011)
Universidad Carlos III de Madrid

Contenido

Tabla de ilustraciones

Índice de tablas

Abstract

Introduction

A study about the Twitter social network will be presented in this document, more specifically we will analyze those cases where the users can become hostile or present rivalry against other users.

For this propose, we will analyze tweets that have been published around a context, due to social network are a diary communication tool, we will study different cases where Twitter's users speak and show their opinion about a topic and how is the sentiment of this opinions.

As examples, we are going to use contests in which users vote or prefer an artist, group or specific person, or situations that could present rivalry between their participants, like football matches or politic debates and in this way we will analyze the content of the tweets in order to classify them like positive or negative to then draw more global conclusions about groups of people that use Twitter, to determinate possible cases of rivalry or hostility between this platform's users.

Motivation

It is a fact the considerable influence that social network has in almost all our daily life. It became practically a routine to show to the world what we are doing throw social net. On Twitter, we express opinions, feelings, etc. about different topics, people or product and services.

I have accomplished my project around online social networks (hereafter social networks) because I think that it is a very relevant topic from which you can obtain a lot of information of a huge variety categories, we can use social networks for analyze one person's likes and in that way, offer a certain type of product, or analyze their professional profile for recruit employees of a company and so on.

For this reason, I also found interesting to make a sentiment analysis of this social network, to understand better all the information that we can obtain and that ourselves publish on it.

State of the art

Context

Today, all of our lives are linked to technology, since we start our day we are using electronic devices, such as television, home appliances, computers, the Internet, and our smartphones, which we use for everything.

Focus on this necessity of being connected always, social networks emerge. We have found a new way of expressing ourselves, being connected with our friends and also find new friendships. We can share different contents or watch those that are of our same interest, search information like news or publications of topics that we like and even social networks are an advertising medium for many companies. Definitively, social networks are present in our diary life.

Analysis of social network

Nowadays almost everybody has a profile in a social network. The utility of the profile could be very varied, since create a profile for spend free time to profiles more professional and commercial.

Whatever it's your goal in a social network, each user is going to generate a big amount of information, and it is here where the role of analyze networks takes part.

What is social network analysis?

Social network analysis is defined like "a methodologic and theoretical approximation that emphasizes the study of relations between actors, both relations between people, organizations, countries or things. [3]

A network is composed by nodes, that represent actors, and lines, that represent links, and the objective is to analyze the cohesion between them, the subsets formed, node's centrality and its composition, the relevance that they have in the creation of links…etc.

Thus, the analysis of a network consists in investigate all the relations and links that compose it, due to when we speak about social network we refer to a set of individuals that share some links that connect them, like one specific topic, friendship, relationship, work activity, etc.

Why analyze network?

A lot of people and companies use social network analysis as a tool, we are going to summarize some of its utilities.

Companies and business

In order to change their business philosophy, companies attend to Social Media, in that way they know better the interest and requirements of their stakeholders and become more agile companies that obtain better results. Some of the advantages that they achieve with social network analysis are the improvement of the access to information, the reduction of communication cost and access to innovation catalysts.

In fact, the social networking has become in other interaction channel between clients and companies, and because of that, it is very important for companies know how to analyze them in order to achieve planted objectives.

Detection of frauds and swindle

We can use social networking for detect frauds, like for example insurance companies that detect swindle cases in their customers when they simulate a illness or incapacity to demand a compensation, but companies catch out them when the customers share photos or videos in their Facebook showing a good health condition.

Harassment and buying detection

New technologies have unleashed occurrence of new terms like cyber-harassment or cyber-buying, that is an extension of buying in technologic media, via mobile phones or internet, where the stalker send threatening and blackmailer messages with the aim of undermine the other person's self-esteem.

Thus, technology has caused this new type of harassment, in which the stalker does not have to face to face with the other person and it is hide behind a screen, that also gives impunity to stalker's acts.

It is therefore that social network analysis results very useful in order to detect cases like this, thus being able to curb situations that could become worse.

Goals

The aim of this project is to analyze a set of tweets, all of them extracted from the same context; ergo they speak about a specific topic, and extract the measure of rivalry and hostility that present some users against others.

In fact, we can specify two main goals in our project:

- Obtain a set of tweets organize around a specific topic, we achieve this with a hashtag that define a study community, and later we will create a database with them.
- When this system has been deployed, we will analyze the subsets of tweets that we consider, first of all to determinate the sentiment of the tweets, could be positive or negative and after we make more global conclusions about these subsets.

We are going to following these steps in order to achieve our goals in the project:

- Selection of a hashtag and study community.
- Analysis sentiment preparation.
- Selection of measures and metrics.
- Data mining.
- Data cleaning and enrichment.
- Calculation of values.
- Analysis and discussion of the results obtained.

System analysis

Due to the big amount of data and information that daily circulates on the internet and more specifically in social networking, it is considered to make a research in order to detect rivalry and hostility cases in different scopes of Twitter social network.

Daily huge amount of social data are produced and them grow up constantly, this make more difficult control and analyze that date and appears the necessity of find tools and programs that could make a sentiment analysis, detect the most popular topics in a community and search behavior patterns in communities.

As we mentioned before, the aim of our project is a study about a specific community, which is defined by a hashtag on Twitter and that generate conflicts or discussion topics between their users.

In that way, security is a very critical point in social network analysis and it is here where terms cyber-rivalry and hostility make sense because if we accomplish detect violent

and unappropriated behaviors, attacks or aggressions, we could create a more security environment in social networking.

To understand better our analysis, is necessary to define both concepts rivalry and hostility:

1. Rivalry [10]: The rivalry is an argument established between two or more individuals, or groups, that follows achieve the same purpose, or the way for accomplished it. Because of that, the decision that provokes a person to position himself as a rival of other person has a list of objectives to satisfy, and a justification to this attitude.

2. Hostility [11]: Is the quality of hostile, that means a provocative and opposite attitude, normally without an apparent reason, against other human being. Hostility involves an abusive and aggressive manner that can be reflected in emotional or physical violence, provoked by only one person, a small group or a big one, and it is lead to one or more subjects.

System Design

In this part, we are going to summarize the steps that we follow to design and implant our system.

The first step is obtain the data, for this we choose Twitter's API because it allows to access to a big amount of public information such as tweets generated by the users. Nevertheless we found some limitations like that we need actual hashtag to recover his tweets, because API recovers tweets in real time, so if we want pass tweets the script that we are using will make an error.

We use a Python script for obtain the data, and in order to his working, it obtains a list of tweets that will be extract in JSON format to a specific folder. Like parameters to introduce, we have the hashtag that will contain all the tweets extracted and the number of tweets that we want to recover.

The second step is store the obtained data, the idea is store them in our own database using ArangoDB, which allow us create documents' collections.

We decided to create collections that store all the tweets that we recover with a common hashtag, in order to create them we use the graphic interface that has ArangoDB and the Shell. We add the documents to the collections using a command that allows us to import a big amount of data in JSON format.

When we have store the data in our ArangoDb database, it is time to treat them in order to obtain their polarity. For this objective, we are going to build our own feeling

evaluation engine using and external tool called MonkeyLearn. With this tool, we are able to build a learning model with which we will analyze the feeling of the tweets.

To create the learning model we need to make a module, that it is like a project, and insert training examples. With the examples, the tool will make a category tree; in our case, the categories will be positive and negative. The set of examples are a group of tweets classifieds previously like positives and negatives and with them MonkeyLearn will be able to obtain a reliable model.

Thus, we will use the model to classify our corpus of tweets. First, we have to clean the data and we use AQL in order to make a query and extract the important data because not all the data stored will be useful. With the query, we will obtain the user that writes the tweet, the text that we want to classify, the hashtags and mentions, and the tweets which language is English.

Finally, we will use the text classify of the MonkeyLearn model created, and we will obtain an excel table with the data explained previously and the classification as positive or negative with a percentage of probability.

On the other hand, we are going to use a formula for get the new measure that we will determine like cyber-rivalry and hostility. First of all we need to explain the following concepts:

- Twitter's users generate resources in a community.
- A community on Twitter will be defined by a list of elements that compose it, like for example a common hashtag or key words that refer to a topic or specific resource.
- Communities can refer to each other and build relationships between them, but we cannot infer on the polarity of this relation because we will need more attributes that help us to study and understand why this communities are partitioned. Because of that, we will establish different cases with respect to relations between members of the same community, depending on the attributes and references that we have.
- In fact, our formula will apply the sentiment analysis for detect the polarity and see how this influence when we have to measure the degree of antipathy between communities.

On that way, being u_i a user of a community C_i that generates a list of sources R for this community, this subset will has his own descriptors for this community C (for example: hashtag) but can contains references or allusions to other communities (hashtags that references other groups or mentions).

We will calculate the rivalry as follows:

$$R(C_i, C_k) = \text{abs}\left(\frac{\sum_{i=1}^{n} S(\, r_{C_i}^{i}, F_{C_k}^{descriptors})}{\mathrm{N}^{C_i} + \mathrm{M}^{F_{C_k}^{descriptors}}} - 1\right)$$

Equation 1 Formula for Cyber-rivalry

- $r_{C_i}^{i}$ it is a resource (tweet) generated inside the community C_i
- -$F_{C_k}^{descriptors}$ represents the set of descriptors of the community C_k
- The function $S(\, r_{C_i}^{i}, F_{C_k}^{descriptors})$ analyzes the sentiment and measure the polarity between the source $r_{C_i}^{i}$ and $F_{C_k}^{descriptors}$
- - N^{C_i} and $\mathrm{M}^{F_{C_k}^{descriptors}}$ represents the cardinality of the resources R_{C_i} and $F_{C_k}^{descriptors}$

The result of the equation will be a value between 0 and 1, where 0 means that the rivalry is null and 1 determine that the communities are totally antagonistic.

For calculate the hostility we will use the following equation:

$$H(C_i) = \sum_{j=1}^{n} R\left(C_i, C_j\right) = \text{abs}\left(\frac{\sum_{i=1}^{n} S(\, community's\ tweets)}{N(\, total\ tweets\,)} - 1\right)$$

Equation 2 Formula for hostility

Which result will be the total sum of the rivalry calculates for the subsets that composes a global community, in that way, we will calculate the function S for all the occurrences that composes the community and with this we will calculate the total hostility of the subset.

Results and experimentation

When we have all the data in an Excel table with the fields that we consider useful, we must do some work to enrich and clean them.

The first that we do was delete needless fields and which are store on the database, with the aim of clean our Excel and only have the main data.

On the other hand, we made a clean of some tweets that contained words in non-Latin alphabets, although we filtrated the data by language, in some cases appear characters of other alphabets.

For the definition of community to study, we explain that we will use a common hashtag that will appear in all posted tweets by the different users, nevertheless, we also need disintegrate the own community in order to find two potentially rivals groups.

This job has presented certain limitations and difficulties, because is complex to find the way to know if one user is referring specifically to other of an opposite community.

Therefore, we took the decision of explain different cases to define communities inside other more global and thus can measure its rivalry and hostility.

- ✓ Case 1: Two different communities appear in the same tweet.
  It will happen when we find users that will tweet a text in which appear mentions or hashtags that refer to two communities. With this we can deduce that the user is follower of one of the two communities that mentions, and reference to a second, so we can try to measure if exists rivalry between both mentioned communities.
- ✓ Case 2: Global community and mention of sub community.
  For this case, we will find that users mention to other community in their tweets, this community is different of the global one.
- ✓ Case 3: Global community
  We choose all the community differenced by her hashtag for do the calculus of the hostility. We will apply this option for cluster all the sub communities inside the big one

Conclusions and future work

Conclusions

To conclude, we think that we reach the objectives initially proposed in the project. We get the measure of rivalry and hostility between different Twitter communities by a sentiment analysis.

Like critical parts of the project, we can focus on the data collection on Twitter, apparently, it looks like an easy task because of the big amount of information that we can find on the net. Nevertheless, in fact, it was not easy to find the particular data with which we can work, due to it must to meet a number of requirements for reach a successful analysis. These requirements are the language or find descriptors that differentiate communities inside Twitter.

Is also an important aspect the sentiment analysis of the tweets, because of it is very important in order to apply our rivalry and hostility equations and they will determine clearly the result of this equation. Although there are many tools that can classify text by their sentiment, we took the decision of use a tool that allows us to create our own prediction model in order to get more accurate data and which were more in line with our classification objective.

Respect to the obtained results, conclusions have been drawn about the rivalry communities, like which are the social groups that presents more conflicts, like the case of football or politics. In addition, we were able to analyze the factors that take part when we have to measure the antipathy between groups, this factors are for example the popularity on Twitter or if they are winners in a contest.

Furthermore, during the develop of the project knowledge acquired at the degree has been put into practice, such as project life cycle and its different phases since the planning to the execution and the presentation of the final result.

Personally, all the tools that make possible data analysis in social network have surprised me, and all the utilities and applications that it has. Specially the big amount of information that without almost realizing, we expose daily at the web and of which we can extract many conclusions, behavior patterns, topic interest, our likes or by the opposite side, the things that disgusting us.

Future work

Some improvements are proposed like addition to the accomplished project:

- Improvement of the prediction model for a more accurate classification : we might develop a prediction model with more training examples, besides attune the tweets that we use as example, classifying them with more details and deleting words and empty characters that do not influence at the moment of determine the sentiment.

  With regard to precision and recall measures, it would be to approximate these values all as we be able to 100% so that the accuracy of the model will also be around this percentage. With this improvement, we might obtain a very reliable model and with little bit errors in order to classify the information.
- To add more factors that determine the relation between two communities. In our study, we detect that one community have relation with another through the used hashtags, nevertheless, it would be a big advance can determinate how two communities are link together, analyzing patterns that they may determine that certain users are referring specifically to other groups in their tweets.
- Creation of a Spanish classifier to be able to do studies about communities that are more diverse and not only the English ones. The difficult of find content already classified in other language different of English, makes that our projects focus on English speaker communities because is the only way of classify the tweets. Like improvement, we could add a Spanish corpus and classify them like positive or negative in order to add them as training examples and then get a Spanish evaluation engine.

1. Introducción

En el presente documento se desarrollará un estudio sobre la red social Twitter, concretamente se analizan los casos en los que los usuarios pueden llegar a ser hostiles o presentar cierta rivalidad con respecto a otros usuarios.

Para ello se analizarán los tweets que se han publicado en torno a un contexto, dado que las redes sociales son una herramienta más de comunicación diaria, se estudian diferentes casos en los que los usuarios de Twitter hablan y opinan sobre un tema y cómo es el sentimiento de esas opiniones.

Como ejemplos se utilizan concursos actuales en los que los usuarios voten o tengan preferencia por un artista, grupo o persona concreta, o escenarios que puedan presentar cierta rivalidad entre sus participantes, como partidos de fútbol o debates políticos y de esta manera se analiza el contenido de sus tweets para clasificarlo como negativo o positivo para después sacar conclusiones más globales acerca de los grupos de personas que utilizan Twitter, para determinar posibles casos de rivalidad u hostilidad entre los usuarios de esta plataforma.

1.1. Situación-Contexto

Actualmente prácticamente toda nuestra vida está ligada a la tecnología, desde que nos levantamos hacemos un uso constante de los dispositivos electrónicos, desde la televisión, los electrodomésticos, el ordenador e Internet, hasta nuestros inseparables móviles smartphones, los cuales utilizamos para prácticamente todo.

Centrándonos en esta necesidad de estar siempre conectados en la red, aparecen las redes sociales, con las que hemos encontrado una nueva manera de expresarnos, de conectarnos con nuestros amigos o incluso conseguir nuevas amistades, podemos compartir distintos contenidos o ver aquellos que resulten de nuestro mismo interés, buscar información como noticias o publicaciones de temas que nos gusten, e incluso son un medio de publicidad para muchas empresas. En definitiva, las redes sociales están presentes en nuestra vida diaria.

Por otro lado, existen varios tipos de redes sociales, se podrían clasificar en [1]:

- Redes sociales horizontales: Aquellas que no funcionan en torno a un tipo específico de usuario o un tema o tópico concreto. Permiten la libre participación de quien lo desee, son ejemplos de este tipo Facebook o Twitter.
- Redes sociales verticales: Este tipo de redes sí que está dedicada a un tipo de usuario concreto, son especializadas y las personas que agrupan tienen un interés común. Es el ejemplo de LinkedIn, que agrupa perfiles profesionales con

el fin de originar relaciones laborales, o redes que agrupan perfiles de ocio con intereses compartidos como deportes, música o videojuegos.

Mi proyecto se desarrolla en torno a la red social Twitter, si hablamos un poco de su historia esta red fue creada en el año 2006 y fue una idea surgida de un proyecto de investigación dentro de una pequeña compañía [2].

Su idea principal destacaba por su sencillez, Twitter sería una red en la que los usuarios se comunicarían entre ellos mediante 140 caracteres, y actualmente tras todo el éxito que ha cosechado, sus cambios han sido prácticamente nulos, exceptuando las distintas mejoras, la idea sigue siendo la misma que en sus comienzos.

Una de las cosas que sí ha cambiado ha sido la publicidad, ya que hoy Twitter permite la promoción de cuentas y tweets mediante un sistema con el que las empresas realizan un pago y pueden anunciarse durante un tiempo, las cuentas promocionadas o los tweets "Promoted" hacen referencia a cuentas que las personas no siguen actualmente pero les pueden resultar interesantes, son muy utilizadas para aumentar el número de seguidores con fines de publicitar empresas.

Una de las claves de Twitter son los Hashtags, mediante el símbolo # seguido de una o varias palabras sin espacios se crean esta especie de etiquetas con las que podemos realizar un seguimiento de temas, con ellos se consigue que los usuarios lleguen de una forma más rápida a ciertos temas y que estos estén organizados.

También se debe mencionar la inclusión de los *Trending Topics,* los temas del momento, su cometido es resaltar aquellos temas que más se repiten en los tweets en un determinado tiempo. La mayoría de los Trending Topics surgen de los hashtags, se comienza a hablar sobre algo en un tweet que contiene un determinado hashtag para referirnos a ese tema, y cuando "todo el mundo" comienza a hablar de ello esos temas se clasifican como Trending Topics y aparecen en un apartado para que los usuarios puedan reconocer o seguir los temas más actuales.

Ilustración 1: Trending Topics

Por lo tanto, Twitter es una plataforma en la que los usuarios pueden expresar su opinión sobre algún tema, comentar sus vivencias diarias, acompañar sus publicaciones de fotos, vídeos o enlaces, y compartirlo con otros usuarios.

1.2. Motivación

Es un hecho la gran influencia que presentan las redes sociales en casi todos los ámbitos de nuestra vida cotidiana. Se ha vuelto prácticamente una rutina mostrar al mundo lo que estamos haciendo cada día a través de ellas.

En Twitter expresamos en unas líneas desde una vivencia de nuestro día hasta nuestra opinión sobre un tema.

He realizado mi proyecto en torno a las redes sociales porque me parece que es un tema muy actual del que se puede sacar mucha información de categorías muy variadas, podemos utilizar las redes para analizar los gustos de una persona y así ofrecer un determinado tipo de producto, o analizar su perfil profesional para captar trabajadores de una empresa y un largo etcétera.

Por este motivo me ha parecido interesante realizar un análisis de esta red social, para así poder comprender mejor toda la información que se puede obtener y que nosotros mismos publicamos en ella.

1.3. Objetivos

El objetivo de este proyecto es analizar un conjunto de tweets, todos ellos extraídos de un mismo contexto, es decir tratan un tema concreto, no son tweets elegidos al azar, y de ellos extraer el grado de rivalidad u hostilidad que presentan ciertos usuarios con respecto a otros.

Por lo tanto, podríamos concretar dos objetivos principales en nuestro trabajo:

- Obtener un conjunto de tweets organizados en torno a un tema concreto, esto lo conseguiremos mediante un hashtag que define una comunidad a estudiar, y posteriormente los almacenaremos creando una base de datos con los mismos.

- Una vez montado este sistema, analizaremos los subconjuntos de tweets que creamos convenientes para primero determinar el sentimiento de dichos tweets, pudiendo ser positivo o negativo y posteriormente sacaremos conclusiones más globales aplicadas a esos subconjuntos.

2. Estado del arte

En este apartado se expondrán los ámbitos en los que se encuadra el trabajo. Para ello se hablará de las redes sociales, cuáles son las que más se usan en todo el mundo y además se expondrán algunas herramientas para su análisis y se detallará la importancia y utilidades de analizar las mismas para así comprender mejor el objetivo del trabajo.

2.1. Análisis de redes sociales

Actualmente casi cualquier persona posee un perfil en alguna red social, la utilidad del mismo puede ser muy variada, desde crear un perfil por mero entretenimiento a perfiles más profesionales y comerciales.

Sea cual sea tu objetivo al entrar en una red social, cada usuario va a generar una gran cantidad de información o va a buscarla, y es aquí donde el papel de analizar las redes toma importancia.

¿Qué es en análisis de las redes sociales?

El ARS se define como "una aproximación metodológica y teórica que enfatiza el estudio de las relaciones entre actores, tanto relaciones entre personas, organizaciones, países o cosas" [3]

Las redes se componen de nodos, que representan los actores, y líneas, que serían los enlaces, y el objetivo es analizar la cohesión entre ellos, los subgrupos formados, la centralidad de los nodos y su composición, la relevancia que tienen en la creación de enlaces.etc.

Por lo tanto, analizar una red consiste en investigar todas las relaciones y enlaces que la componen, puesto que cuando hablamos de red social nos referimos a un conjunto de individuos que comparten ciertos vínculos que les unen, ya sea un tema de interés concreto, amistad, parentesco, actividad laboral…etc.

Principales redes sociales

Actualmente existe un gran número de redes sociales, como introducción se explicarán brevemente algunas de las más importantes a nivel mundial según el ranking publicado en WebEmpresa2.0. En esta lista, Twitter aparece en la posición número 11 y puesto que este trabajo se centra en esta red, se le dedicará un apartado más adelante.

Facebook

Clasificada como la red más popular del mundo, Facebook fue creada por Mark Zuckerberg en el año 2004. Lo que empezó como un pequeño proyecto creado por estudiantes de la universidad de Harvard, acabó convirtiéndose en una red que tiene alrededor de 1.508 millones de usuarios registrados alrededor de todo el mundo [4]

La idea inicial que motivó Facebook, fue la creación de un espacio en el que los alumnos de la universidad pudieran comunicarse e intercambiar información, comenzaron creando un directorio con las fotos de los anuarios para que lo utilizaran las fraternidades de la universidad.

La red tuvo un éxito abrumador, y creció exponencialmente hasta convertirse en la red que actualmente permite que millones de personas compartan fotos, actualizaciones de estado, vídeos, contenidos de otras webs y sobre todo conecta a multitud de personas entre sí.

YouTube

En el segundo puesto de la clasificación, aparece YouTube cuya principal funcionalidad es la de ver, subir y compartir vídeos.

Esta red fue fundada en el año 2005 por tres exempleados de PayPal [5], y sus orígenes los encontramos en una fiesta, dos de los fundadores grabaron un vídeo y al querer enviarlo por correo electrónico se encontraron con que era demasiado largo para poderlo enviar, de manera que tuvieron la idea de crear un sitio web en el que cualquier usuario pudiera enviar y ver vídeos.

Actualmente se cuelgan unos 65.000 vídeos al día y en octubre de 2006 YouTube es comprado por Google por 1.650 millones de dólares**.**

2.2. Herramientas para el análisis de redes sociales

En este punto se detallarán algunas herramientas que realizan un trabajo similar al que va a realizar el presente sistema. Nos centraremos en aquellas que realicen un estudio de la red social Twitter ya que en ella se centra este trabajo.

Twitter Analytics

Es una herramienta gratuita que permite el análisis de contenido de Twitter. Muestra una gran variedad de datos de manera sencilla para que podamos conocer y analizar nuestra presencia en Twitter, tanto para saber que demandan nuestros seguidores o cómo reaccionan estos ante nuestras publicaciones [6]

Se centra sobre todo en una representación visual de los datos mediante su interfaz gráfica, en la que nos muestra los diferentes números y porcentajes de los datos relevantes en nuestra cuenta. Entre estos datos podemos encontrar:

- Mención principal: Contiene la publicación en la que hemos sido mencionados con mayor número de interacciones durante el mes.
- Tweet principal: Aparece el Tweet que más impresiones ha obtenido en el mes, lo que es útil para saber que le gusta a nuestra comunidad de seguidores y potenciarlo.
- Tweets: La herramienta cuenta con una pestaña en la que aparecen datos a tiempo real sobre la interacción que conseguimos con nuestras publicaciones, además muestra su evolución durante los últimos 28 días

En definitiva, con esta herramienta se puede analizar los datos de Twitter para saber qué les gusta a nuestros seguidores, cuál es la mejor hora para realizar una publicación y que tenga éxito, datos sobre nuestros seguidores como datos demográficos o estilo de vida para destacar intereses y eventos que se realizarán en nuestro país y que serán útiles para analizar su repercusión en Twitter.

Ilustración 2 : Twitter Analytics [6]

Twitter Sentiment

Se hablará ahora de algunas herramientas que se centran en el análisis de sentimiento de los tweets y por tanto están estrechamente relacionadas con la herramienta desarrollada.

La primera es Twitter Sentiment [7], es básicamente un buscador en el que introducimos el término del cual queremos consultar los sentimientos de los tweets, como por ejemplo una marca o producto.

Después de buscarlos, la herramienta mostrará una representación gráfica de la evolución de tweets positivos y negativos, un análisis de la tendencia y la popularidad y un listado de los últimos tweets que se han generado.

2.3. ¿Por qué analizar las redes?

El tema de análisis de redes sociales está a la orden del día y cada vez más personas y empresas lo utilizan como herramienta, aquí se explican algunas de las utilidades que se le da a dicho análisis.

EMPRESAS Y NEGOCIOS

Las nuevas tecnologías y en concreto las redes sociales, están cambiando la manera tradicional de realizar muchas cosas, en este caso han modificado las tradicionales prácticas empresariales que conocemos [8].

Las empresas acuden al Social Media para transformar su filosofía de negocio, de esta forma conocen mejor los intereses y necesidades de sus stakeholders y se convierten en empresas más ágiles que obtienen mejores resultados. Algunas de las ventajas que experimentan con el uso del análisis de las redes son la mejora al acceso de la información, la reducción en los costos en comunicación y el acceso a catalizadores de innovación.

Para las empresas el análisis web surge con la finalidad de optimizar la estrategia digital y así sacar mayor rentabilidad al dinero invertido. Por lo tanto, el principal objetivo del análisis de las redes en las empresas es económico, ya que eligen cuidadosamente qué es lo que necesitan medir para así conocer la estrategia que mejor funcionará para reducir lo máximo posible los gastos y conseguir resultados esperados.

Por lo tanto, las redes sociales se han convertido en otro canal más de interacción entre clientes y empresas, y por ello es de gran importancia para estas últimas saber analizar las mismas en base a lograr los objetivos planteados.

DETECCIÓN DE FRAUDES Y ESTAFAS

Otra de las utilidades que tiene analizar las redes sociales, es la detección de posibles fraudes y estafas de todo tipo.

Como ejemplo tenemos a las compañías aseguradoras que han detectado casos de estafa en sus asegurados, para ello no hace falta realizar un exhaustivo análisis si no que muchas veces son los propios usuarios los que se delatan. Nos encontramos con casos en los que personas fingen incapacidades, por ejemplo, no pueden andar, y exigen una indemnización a sus compañías, pero estas tras analizar redes como por ejemplo Facebook, pueden encontrar vídeos o fotografías que estas personas han colgado bailando o andando y que les delatan sin ninguna duda.

DETECCIÓN DE ACOSO Y BULLYING

El uso de las redes sociales se da principalmente en grupos de adolescentes y jóvenes, ya que son la generación que ha nacido y crecido con la tecnología.

Es por ello que analizando los contenidos que publican podemos conseguir detectar el acoso que tantos niños y jóvenes sufren actualmente.

En artículos como [9] se describe un nuevo término, el ciberacoso o ciberbullying, que es una extensión del acoso en los medios tecnológicos, teléfono o Internet, y mediante el cual el acosador envía mensajes amenazantes o chantajistas con la finalidad de minar la autoestima de la otra persona. El medio que utilizan para enviar estos mensajes suele ser tanto e-mail, mensajes de texto, mensajería instantánea o las redes sociales.

Por lo tanto, la tecnología ha desencadenado la aparición de este nuevo tipo de acoso, en el que el acosador no tiene que enfrentarse cara a cara con la persona y se esconde tras una pantalla, lo que además da al acosador cierta impunidad en sus actos.

Es por ello que el análisis de las redes resulta de gran utilidad a la hora de detectar casos como estos, pudiendo así poner freno a situaciones que podrían tornarse a peor.

3. Análisis del sistema

Debido a la gran cantidad de datos e información que diariamente circula por Internet y más concretamente en las redes sociales, se plantea realizar un estudio para detectar casos de rivalidad u hostilidad en distintos ámbitos de la red social Twitter.

Diariamente se producen grandes cantidades de datos sociales y crecen constantemente, esto hace más complicado su control y análisis y surge la necesidad de buscar herramientas que puedan clasificar este contenido. Se han ido desarrollando herramientas y programas que analizan el sentimiento, detectan los temas más hablados en una comunidad o buscan patrones de comportamiento en ellas.

En este contexto, la seguridad sigue siendo un punto muy crítico en el análisis de redes y es ahí donde entra en juego los términos cyber-rivalidad y hostilidad, ya que, si conseguimos detectar comportamientos violentos o agresivos, ataques y agresiones, podremos crear un entorno más seguro en las redes.

Como se ha comentado antes, el objetivo de este trabajo es el estudio de una comunidad concreta definida por un hashtag en Twitter que genere conflictos o temas de discusión entre sus usuarios. Se añade un esquema conceptual del diseño del mismo:

Para comprender mejor dicho análisis, es necesario definir los conceptos de rivalidad y hostilidad:

1.Rivalidad [10]: La rivalidad es la disputa que se establece entre dos o más individuos, o entre grupos, a la hora de conseguir un mismo fin, o en el camino a superar el mismo. Por esto, la decisión que suscita en una persona posicionarse como rival de otro, tiene una serie de objetivos a cumplir, y una justificación a tal actitud.

2.Hostilidad [11]: es la cualidad de hostil, que indica una actitud provocativa y contraria, generalmente sin motivo alguno, hacia otro ser vivo. La hostilidad, por lo tanto, implica una conducta abusiva y agresiva que puede reflejarse en violencia emocional o física, de mano de una sola persona, un grupo pequeño o una gran cantidad de gente y estar dirigida, de igual forma, a uno o más sujetos.

3.1. Fuentes a utilizar

A continuación, se describirán las fuentes de datos que han sido utilizadas para recabar la información que necesitamos para nuestro análisis.

Twitter

Es la red social en torno a la que se desarrolla este trabajo. Como ya se ha comentado se trata de una aplicación que no ha parado de crecer desde su creación y actualmente cuenta con unos 1,3 mil millones de cuentas creadas. Algunos datos curiosos son [12]:

- Existen 310 millones de usuarios activos al mes.
- Un 29,2% de los usuarios de redes sociales en Estados Unidos son usuarios de Twitter
- El 80% de los usuarios activos acceden a la plataforma a través del móvil.
- El 83% de los líderes mundiales están en Twitter.
- Se mandan 500 millones de tweets al día. Eso quiere decir 6000 tweets por segundo.
- "Lágrimas de risa" es la emoji más usada en Twitter, con 14,5 mil millones de tweets.

3.1.1. API Twitter

Para comenzar... ¿Qué es una API?
Su nombre viene de la abreviatura en inglés de Application Programming Interfaces (Interfaces de programación de aplicaciones)[13] , es una especificación formal , es decir, el conjunto de comandos, funciones y protocolos que permiten a los desarrolladores de software crear programas para ciertos sistemas operativos, simplificando así el trabajo al crear un programa, ya que sirve como "plantilla" para no tener que empezar a escribir un código totalmente desde cero pudiendo usar funciones predefinidas.

Las APIs de Twitter

Twitter ofrece tres APIs en función a diferentes necesidades, Streaming API, REST API y Search API.

Ilustración 3 : Api de Twitter[13]

- Streaming API: El principal cometido de esta API es proporcionar un subconjunto de tweets prácticamente en tiempo real, mediante una conexión permanente por usuario con los servidores de Twitter y realizando una petición http se recibe un flujo de tweets de manera continua, en formato JSON. Además, podemos filtrar los tweets que queremos obtener, pudiendo así conseguir una muestra aleatoria, un filtrado por palabras clave o usuarios, los tweets con enlaces, los tweets con retweets...etc.
- Search API: Con esta API podemos obtener los tweets que se ajustan a una query solicitada con una antigüedad de 7 días. Se pueden aplicar filtros por cliente utilizado, lenguaje y localización. Por otra parte, esta API también muestra información más particular del tweet, como por ejemplo, datos de autor como el Id, screen_name y la url de su avatar.
- Rest API: Orientada a los desarrolladores, permite el acceso al core de los datos de Twitter. Además, posibilita realizar mediante la API las operaciones que se realizan vía web soportando varios formatos como xml, json, rss, atom...

3.1.2. Campos de interés

Consultando la documentación de la API [14] que nos ayudará a comprender mejor su funcionamiento básico, nos centraremos en los campos y objetos de un tweet que serán útiles a la hora de analizar su contenido.

Para comenzar, hay cuatro “objetos” principales que encontraremos en la API, Tweets, Users, Entities y Places. A continuación, se explicarán los que han sido de utilidad.

Tweets

Los tweets son el bloque más básico y en torno a los que se estructura todo en Twitter. Como hemos mencionado, están compuestos de un máximo de 140 caracteres y pueden incluir contenido de distintos tipos, además pueden ser:

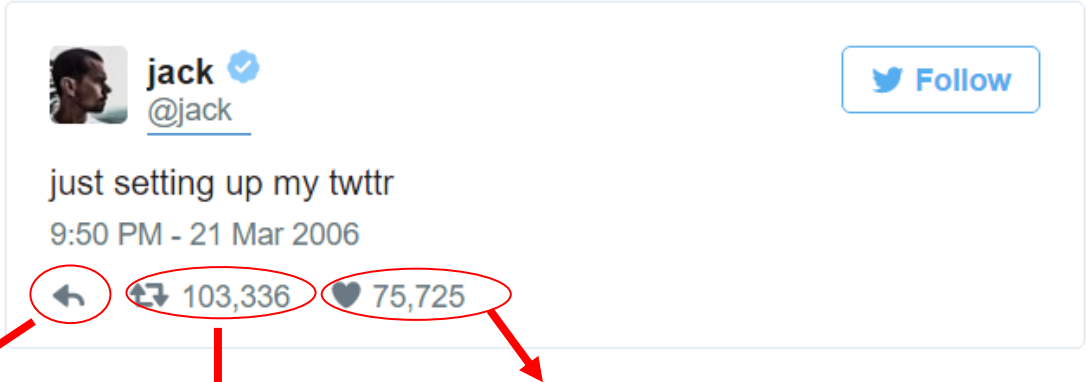
- Embedded: Este tipo de tweets nos permite añadir el contenido creado en Twitter en nuestros artículos o webs. Un embedded Tweet puede incluir fotos o vídeos creados para mostrarse en Twitter, o links interactivos y pre visualizaciones de contenidos adicionales. El significado de embedded es incrustado o encajado, un elemento que se integra dentro de uno más grande, y esta es la funcionalidad de estos tweets, proporcionar un código HTML para insertar el contenido en un blog o Web.
- Replied to: Este tipo de Tweet responde al Tweet de otro usuario y comienza con el @nombredeusuario del usuario al que quiere responder.
- Liked: Un tweet con esta característica es similar a un Like en Facebook. Marcando esta opción tenemos la posibilidad de interactuar con el contenido de cualquiera de los favoritos, mostrar nuestra apreciación o simplemente notificar al autor que hemos visto su Tweet.
- Unliked: Si hemos marcado la opción de favorito en un Tweet podemos deshacer esta operación pulsando otra vez el icono con forma de corazón del Tweet.
- Deleted: Ocurre cuando borramos un Tweet que habíamos publicado previamente.

En la ilustración mostrada a continuación, se puede observar un ejemplo de como se mostraría un Tweet en la aplicación Twitter.

Aparece en nombre del usuario que lo postea, con su avatar, seguido de su alías en la aplicación, que aparece predecido de @.

A continuación el contenido del tweet, con la fecha y la hora en la que fue publicado.

Ilustración 4 Ejemplo de Tweet



Permite responder al tweet

Número de retweets de ese tweet y permite retweetear

Muestra el número de favoritos de ese tweet y permite márcalo como tal

Como podemos observar en la documentación de la API,los tweets están compuestos de varios campos y es común que estos varíen o se añadan nuevos dependiendo del caso, ya que no todos los campos aparecen en todos los contextos. A continuación mostraremos ejemplos de esos campos y nos centraremos en las Entities ya que consideramos de importancia para nuestro trabajo.

Campo	Tipo	Descripción
created_at	String	Tiempo UTC en el que el tweet fue creado
favorite_count	Integer	Indica cuantas veces ha sido liked por un usuario de Twitter
id	Int64	Representa un identificador único para ese tweet
user	User	El usuario que ha posteado ese tweet
retweet_count	Int	Número de veces que ese tweet ha sido retweeteado
text	String	Es el texto del tweet, está codificado en UTF-8 y se utilizan unas "reglas" para definir los caracteres que están permitidos

Tabla 1 Campos de un tweet

Entities

Las entidades proporcionan metadatos e información contextual adicional sobre el contenido posteado en Twitter. Los campos contenidos son los siguientes:

Campo	Tipo	Descripción
Hashtags	Array of Object	Representan los hashtags en el texto del tweet
Media	Array of Object	Elementos multimedia cargados en el tweet
Urls	Array of Object	Representan las URLs incluidas en el texto de un tweet
User_mentions	Array of Object	Representan otros usuarios de Twitter mencionados en el tweet

Tabla 2 Campos de una entidad

Analizaremos con más detalle los campos que nos van a ser útiles:

Hashtags

```
"hashtags":[{"indices":[32,36],"text":"lol"}]
```

Ilustración 5 Ejemplo código de un hashtag

Campo	Tipo	Descripción
Indices	Array of int	Array de enteros que indica dónde comienza y termina el hashtag en un tweet. El primer número indica la localización del carácter # en el string de texto del tweet. El segundo número representa la localización del primer carácter después del hashtag, por lo tanto la diferencia entre ambos números será la longitud del hashtag más uno (para contar el carácter #)
Text	String	Nombre del hashtag menos el carácter #

Tabla 3 Campos de un Hashtag

User_mentions

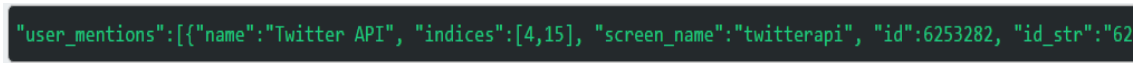


Ilustración 6 Ejemplo de código de una mención de usuario

Campo	Tipo	Descripción
name	String	Muestra el nombre del usuario referenciado
indices	Array of int	Array de enteros que representa dónde comienza y termina la referencia al usuario, sin el texto del Tweet. El primer número indica la localización del carácter '@' de la mención del usuario, el segundo indica la localización del primer carácter después de la mención del usuario.
Screen_name	String	Screen name del usuario referenciado.
id	Int64	ID del usuario mencionado, representado como un entero.
Id_str	String	ID del usuario mencionado, representado como un String

Tabla 4 Campos de una user_mention

Users

Hemos mencionado ciertos campos pertenecientes a este objeto, por lo que los explicaremos brevemente.

Los usuarios pueden ser cualquier persona o cualquier grupo, por ejemplo nosotros seríamos usuarios, pero también lo son varias personas que tienen un usuario para apoyar a un grupo concreto o cuentas publicitarias etc.

Los usuarios tweetean, siguen, crean listas y tienen un timeline de inicio, pueden ser mencionados y buscados.

Ilustración 7 :Ejemplo de usuario

Campo	Tipo	Descripción
id	Int64	Identificador único para este usuario, se representa con un entero y se suele utilizar un entero de 64 bits con signo para que sea seguro.
name	String	El nombre del usuario, tal y como él lo ha definido. No es necesario que sea el nombre real de la persona, normalmente está limitado a 20 caracteres, pero está sujeto a cambios.
screen_name	String	Es el nombre o alias que aparece en la pantalla de perfil del usuario y que lo identifica, son únicos, pero esto puede cambiar por lo que es mejor utilizar id_str como identificador de usuario cuando sea posible. Suele ocupar unos 15 caracteres como máximo.
following	Type	Cuando este campo es true indica que el usuario que está autenticado está siguiendo a este usuario.
followers_count	Int	Número de seguidores que la cuenta tiene actualmente.
lang	String	Código utilizado para declarar el idioma en el que el usuario utiliza la interfaz.

Tabla 5 : Campos de User

3.2. Selección de la comunidad a estudiar

Nos encontramos con una red social, Twitter, que está compuesta por una gran cantidad de usuarios que producen información. Sin embargo, para poder aplicar nuestro análisis se necesita particionar o agrupar a esos usuarios en torno a alguna característica común, y es aquí donde entra en juego el estudio de las comunidades.

El atributo o descriptor que se utilizará para definir una comunidad dentro de la red social serán los hashtags, eligiendo uno común sobre el que estén escribiendo un grupo de personas en tiempo real.

Como se ha visto un hashtag es una etiqueta que ayuda a agrupar los tweets de usuarios que están escribiendo en ese momento sobre un tema relacionado. Por ello se buscará un hashtag que haga referencia a un concurso o tema concreto.

A continuación, se hará una breve descripción de las comunidades obtenidas mediante los hashtags #Eurovision, #Bundesliga y #ITVDebate

Eurovisión

Eurovisión es un concurso musical que se celebra una vez al año en el que un conjunto de países realizan una actuación musical que les representa. Se realiza un festival en una fecha determinada y en un país concreto y el ganador es elegido mediante un sistema de votación en el que se tiene en cuenta el voto del público y el voto de un jurado propio de cada país. Los votantes son los distintos países que deben otorgar 12,10,8, 7… puntos a las actuaciones de los países que prefieran.

La participación de los países no está restringida a los miembros de la de la unión europea y este año 2017 han participado 43 países entre los que se encuentran Alemania, Australia, España, Dinamarca, Francia, Irlanda, Israel, Noruega, Malta, Reino Unido, Rusia y Suiza.

Con el conjunto de tweets generados por este concurso con el hashtag #Eurovision, analizaremos el grado de hostilidad que tienen en conjunto y la posible rivalidad que presentan los países entre ellos.

Bundesliga y liga inglesa de fútbol

Por otro lado, también analizaremos una temática diferente a los concursos, como es el fútbol. En este ámbito se suelen dar muchos casos de rivalidad entre aficiones y por ello vamos a analizar el conjunto de tweets relacionados con la Bundesliga[15], que es una competición de fútbol entre los equipos de la máxima categoría en Alemania y comprende tres divisiones, y la final de la liga inglesa que se disputó entre los equipos Arsenal y Chelsea.

En cuanto a la liga alemana, es similar a otras ligas europeas, pero a diferencia de estas no hay un partido denominado como clásico como puede ocurrir en España, por ejemplo. Sí que se dan encuentros similares a los clásicos y que nos interesa analizar por su rivalidad entre equipos, como son el que enfrenta al Bayern Múnich contra el Werder Bremen o el Schalke contra el Borussia Dortmund.

Con respecto a la liga inglesa analizaremos el último partido que se disputó y en el que se diferencian claramente las dos comunidades enfrentadas, Chelsea y Arsenal.

ITVDebate

Con motivo de las elecciones generales de este año 2017 en reino unido, se realizó un debate televisado que tuvo lugar en Manchester en el que los principales líderes de la oposición debatieron y conversaron entre ellos durante dos horas [16].

Este evento fue seguido en Twitter mediante el hashtag #ITVDebate y los participantes fueron los siguientes:

Tim Farron: Liberal Democrats

Nicola Sturgeon: SNP

Paul Nuttall: UKIP

Caroline Lucas: Green Party

Leanne Wood: Plaid Cymru.

3.3. Obtención de una fórmula

El objetivo es detectar cuando un usuario que es fan, que vota a un grupo o artista o es seguidor de alguna comunidad, es hostil o presenta rivalidad con respecto a otros usuarios.

Para ello se debe encontrar una fórmula con la que se obtenga un cierto valor o ratio que permita determinar si los tweets son o no hostiles. Se debe analizar todos los elementos que componen un tweet para así poder obtener conclusiones, de esta manera, la fórmula será el resultado final de disgregar los diferentes componentes de un tweet, buscando así información desde la unidad más pequeña que lo compone hasta el conjunto de todas ellas.

A continuación, se expone un esquema del proceso seguido para la obtención de la polaridad de un conjunto de tweets para posteriormente poder obtener el grado de rivalidad del mismo.

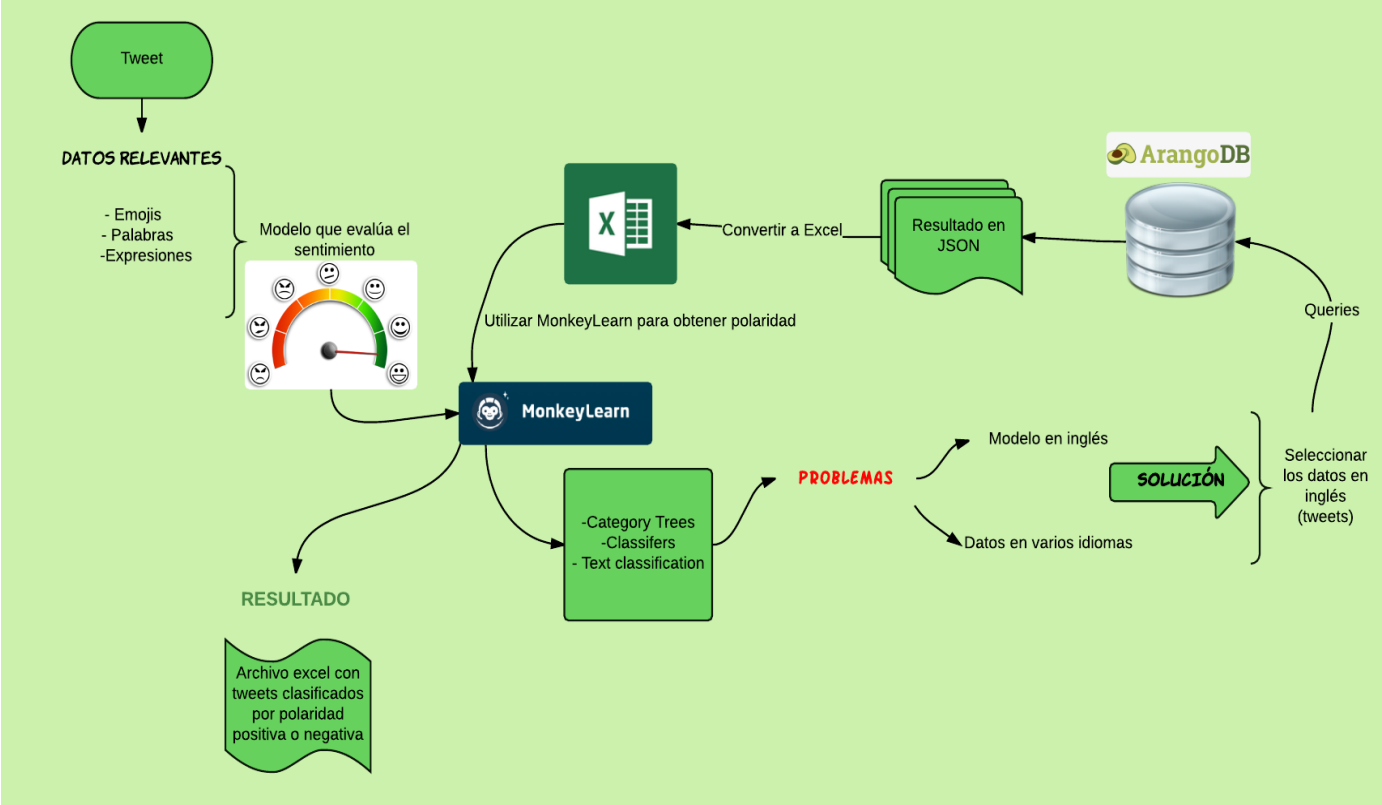


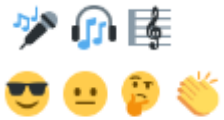
Ilustración 8 : Esquema de análisis de tweets

- Búsqueda de palabras relevantes

El primer paso para obtener el sentimiento de los tweets será comenzar por la unidad más pequeña que lo compone, las palabras.

Calificar las palabras que contenga un tweet como positivas o negativas, de manera que analizaremos el sentimiento de las mismas. Buscar:

- Emoticonos: Se puede obtener una lista con los emoticonos y su significado y evaluar así cuales son negativos y positivos

Posibles emojis positivos	Posibles emojis negativos	Posibles emojis neutros
		

- Palabras: Elaborar una lista con las palabras evaluando su tono.

Posibles palabras positivas	Posibles palabras negativas
Agradable, aceptable, adorar, bueno, bonito, cariñoso, confianza, corazón, divino, dulce, elegante, especial, fácil, fabuloso, gustar, hermoso, incomparable, justo, listo, magnífico, mejor , nuevo , orgulloso , perfecto, precioso , querido , revolucionario , supremo, talento , útil , válido .	Aburrido, avergonzar, basura, confuso, disgustado, enfadado, feo, triste, solitario, molesto, infeliz, amargado, dañino , malo , poco, lamentable, detestar, débil, demonios, soledad, odio, perdedor

- Expresiones: En inglés, abreviaturas como:

Abreviatura	Significado en inglés	Significado en español
LOL	Laughing out Loud	Riendome a carcajadas.
OMG	Oh my GOD	Oh dios mío!
WTF	What the fuck	Para mostrar asombro
LMAO	Laughing my ass off	Para expresar risa
BAE	Before anyone else	Antes que nadie
BFF	Best Friends Forever	Mejores Amigos por Siempre
TGIF	Thank God It's Friday	Gracias a Dios es Viernes
IDK	I Don't Know	No lo sé

Tabla 6: Expresiones en inglés

- Obtención del tono del mensaje

Una vez separado el mensaje en palabras, se puede obtener cuál es la polaridad de ese tweet para posteriormente determinar si estamos ante un usuario hostil o no.

1. Primeramente, se obtiene una diferencia entre las palabras positivas y negativas contenidas:

   Número de palabras positivas - Número de palabras negativas

   Con ello se puede lograr una primera estimación del sentimiento del tweet
2. Análisis del sentimiento:

   En esta parte se debe determinar cuál es la polaridad del mensaje, si se trata de un mensaje positivo por el conjunto de palabras, expresiones e iconos que contiene, o si está ante un mensaje negativo. Para obtener este dato se acude a herramientas que analicen el sentimiento y nos den un resultado óptimo con el que se pueda trabajar.

- Fórmula

Para conseguir la nueva medida de análisis en las redes sociales que se denominará cyber-rivalidad, y posteriormente definir la hostilidad, son necesarias algunas aclaraciones:

- Por un lado, los usuarios de Twitter generan recursos en una comunidad
- Una comunidad en Twitter quedará definida por una serie de elementos que la componen como puede ser un hashtag común o palabras clave que hagan referencia a un tema o recurso concreto.
- Las comunidades pueden referenciarse entre ellas y establecer relaciones entre sí, pero no podemos inferir en la polaridad de esa relación ya que necesitaríamos más atributos que nos ayudaran a estudiar y comprender porque esas comunidades están particionadas. Por ello estableceremos distintos casos con respecto a las relaciones entre miembros de una misma comunidad, dependiendo de los atributos y referencias que tengamos.
- Por lo tanto, nuestra fórmula aplicará en análisis de sentimiento para detectar la polaridad y ver como esta influye a la hora de medir el grado de antipatía entre comunidades.

De esta forma, siendo u_i un usuario de una comunidad C_i que genera una serie de recursos R para esa comunidad, este subconjunto tendrá sus propios descriptores para esta comunidad C (ejemplo: hashtag) pero puede contener referencias o alusiones a otras comunidades (hashtags que referencien a otros grupos o menciones) .Se calculará la rivalidad con la siguiente ecuación:

$$R(C_i, C_k) = \text{abs}\left(\frac{\sum_{i=1}^{n} S(\, r_{C_i}^{i}, F_{C_k}^{descriptors})}{\mathrm{N}^{C_i} + \mathrm{M}^{F_{C_k}^{descriptors}}} - 1\right)$$

Ecuación 1 Fórmula para calcular la Cyber-rivalidad.

- $r_{C_i}^{i}$ es un recurso (tweet) generado dentro de la comunidad C_i

-$F_{C_k}^{descriptors}$ representa el conjunto de descriptores de la comunidad C_k

-La función $S(\, r_{C_i}^{i}, F_{C_k}^{descriptors})$ es la que analiza el sentimiento y mide la polaridad entre el recurso $r_{C_i}^{i}$ y el recurso $F_{C_k}^{descriptors}$

- N^{C_i} y $\mathrm{M}^{F_{C_k}^{descriptors}}$ representan las cardinalidades de los conjuntos R_{C_i} y $F_{C_k}^{descriptors}$

El resultado de la ecuación será un valor entre 0 y 1, dónde 0 indica que la rivalidad es nula y 1 determinará que las comunidades son totalmente antagónicas.

Para el cálculo de la hostilidad utilizaremos la ecuación:

$$H(C_i) = \sum_{j=1}^{n} R\left(C_i, C_j\right) = \text{abs}\left(\frac{\sum_{i=1}^{n} S(\, tweets\ de\ la\ comunidad\,)}{N(\, tweets\ totales)} - 1\right)$$

Ecuación 2 Fórmula para calcular cyber-hostilidad

Cuyo resultado será la suma total de las rivalidades calculadas para las subcomunidades que componen una comunidad global, de manera que, se calcula la función S para todas las ocurrencias que componen la comunidad y con ello se calculará la hostilidad total del conjunto.

Aplicación de las fórmulas

Para comprender mejor las conclusiones obtenidas del análisis realizado anteriormente se propone un caso de ejemplo.

Se parte del siguiente escenario, se obtienen los diferentes tweets publicados por varios usuarios, estos tweets estarán contenidos dentro de una comunidad específica, es decir contendrán un determinado hashtag con el que los clasificaremos. Por ejemplo, utilizando los tweets obtenidos en la disputa de un partido de la final de la liga inglesa entre los equipos Arsenal y Chelsea, se obtendrán los tweets de los usuarios que han utilizado un hashtag para mencionar a un equipo, y de ahí se evaluará su tono.

Tweets	Usuario	Hashtag
Conjunto1	@USUARIO	#Arsenal
Conjunto2	@USUARIO2	#Chelsea

Clasificar los tweets de cada conjunto: Donde ✓ significa que el tweet ha sido determinado como positivo y ✕ como negativo.

	Tweet1	Tweet2	Tweet3	Tweet4	Tweet5	Tweet6	Tweetn+1
Conjunto 1	✓	✕	✓	✕	✓	✓	✓
Conjunto 2	✕	✓	✓	✕	✓	✕	✕

	$F^{d1}_{C_{Arsenal}}$	$F^{d2}_{C_{Arsenal}}$
$r^1_{C_{Chelsea}}$	0,4	0,6
$r^2_{C_{Chelsea}}$	0	0,8

Resultado de la fórmula S de evaluación del sentimiento

$$R(C_{Chelsea}, C_{Arsenal}) = \text{abs}\left(\frac{0{,}4 + 0 + 0{,}6 + 0{,}8}{2 + 2} - 1\right)$$

3.4. Requisitos

En este punto se detallará qué es necesario para que el sistema funcione y cumpla con las expectativas propuestas, para ello se obtendrán los requisitos que este debe seguir.

Debido a que el sistema no es de gran complejidad en cuanto a su función y utilidad, se explicarán los requisitos obtenidos de manera sencilla e informal.

Requisitos Funcionales

Los requisitos funcionales son aquellos que se deben cumplir para que el sistema funcione correctamente, son los más básicos y principales ya que definen la funcionalidad que el sistema pretende ofrecer al usuario.

Código	RF-001
Descripción	El sistema deberá permitir la extracción de conjunto de tweets en tiempo real.
Proceso	Obtención de datos
Prioridad	Alta

Tabla 7: RF-001

Código	RF-002
Descripción	El sistema permitirá la elección de un hashtag como parámetro para la obtención de los tweets
Proceso	Obtención de datos
Prioridad	Alta

Tabla 8: RF-002

Código	RF-003
Descripción	El sistema permitirá elegir el número concreto de tweets que quieren ser recuperados
Proceso	Obtención de datos
Prioridad	Media

Tabla 9:RF-003

Código	RF-004
Descripción	El sistema almacenará los tweets en formato JSON especificado
Proceso	Limpieza y almacenamiento
Prioridad	Media

Tabla 10:RF-004

Código	RF-005
Descripción	El sistema deberá permitir la obtención y recuperación de datos concretos de los tweets almacenados
Proceso	Limpieza y almacenamiento
Prioridad	Media

Tabla 11:RF-005

Código	RF-006
Descripción	El sistema permitirá el análisis del sentimiento mediante la información obtenida de los tweets
Proceso	Análisis de sentimiento
Prioridad	Media

Tabla 12: RF-006

Código	RF-007
Descripción	El sistema utilizará un modelo cuya salida serán los tweets clasificados como positivos o negativos
Proceso	Análisis de sentimiento
Prioridad	Media

Tabla 13: RF-007

Código	RF-008
Descripción	El sistema deberá permitir extraer los datos que se requieran en un formato adecuado para su análisis
Proceso	Limpieza y almacenamiento
Prioridad	Media

Tabla 14: RF-008

Código	RF-009
Descripción	El sistema obtendrá un valor comprendido entre 0 y 1 como medida de la rivalidad
Proceso	Obtención de valores rivalidad y hostilidad
Prioridad	Media

Tabla 15: RF-009

Código	RF-010
Descripción	El sistema obtendrá un valor comprendido entre 0 y 1 como medida de la hostilidad
Proceso	Obtención de valores rivalidad y hostilidad
Prioridad	Media

Tabla 16: RF-010

Requisitos No Funcionales

Los requisitos no funcionales son aquellos que se refieren a las características de funcionamiento y que imponen restricciones en el diseño e implementación del sistema.

Código	RN-001
Descripción	La herramienta obtendrá los tweets en un tiempo de respuesta lo más breve posible
Proceso	Obtención de datos
Prioridad	Media

Tabla 17: RN-001

Código	RN-002
Descripción	La herramienta permitirá almacenar el conjunto de tweets de manera clara para el usuario
Proceso	Limpieza y almacenamiento
Prioridad	Media

Tabla 18:RN-002

Código	RN-003
Descripción	La obtención del sentimiento de los tweets se realizará a partir de la comparación con el modelo de entrenamiento proporcionado
Proceso	Análisis de sentimiento
Prioridad	Media

Tabla 19: RN-003

Código	RN-004
Descripción	El idioma tanto del modelo de entrenamiento como de los tweets será en inglés
Proceso	Análisis de sentimiento
Prioridad	Media

Tabla 20: RN-004

3.5. Casos de uso

En este sencillo diagrama se expondrán los casos de uso de nuestra aplicación:

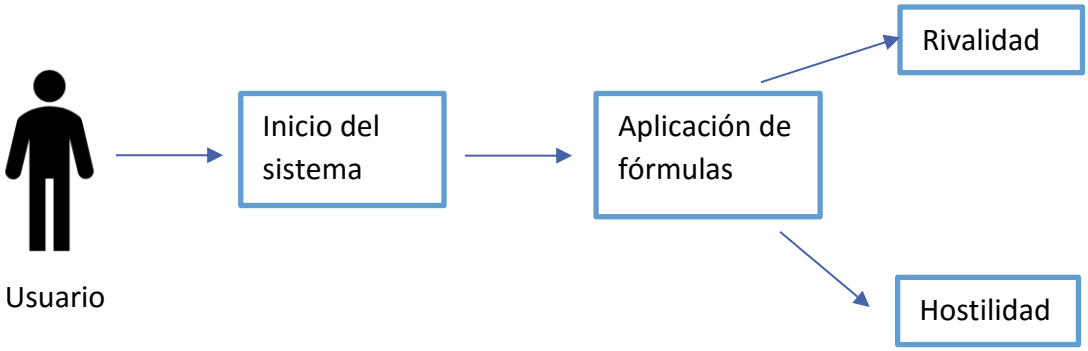


Ilustración 9: Diagrama caso de uso

4. Entorno tecnológico

4.1. Herramientas utilizadas

A continuación se realizará una descripción de las herramientas software y de los lenguajes de programación que han sido utilizados para llevar a cabo el trabajo.

4.1.1. Lenguaje de programación: Python

Python es un lenguaje de programación interpretado y multiparadigma[17], es decir, que permite la orientación a objetos, programación imperativa y en menor medida la programación funcional. Uno de los objetivos de este lenguaje es que el código sea legible, por ello su sintaxis suele ser sencilla favoreciendo así que este lenguaje sea ideal para scripting y desarrollo rápido de aplicaciones en distintas áreas.

Es administrado por la Python Software Foundation. Posee una licencia de código abierto, denominada Python Software Foundation License.

El intérprete de Python y su biblioteca estándar están disponibles de manera libre en el sitio web de Python, en forma binaria y de código fuente para las principales plataformas.

4.1.2. Arango DB

Una vez obtenida la información necesaria, es decir, un conjunto de Tweets que tienen un hashtag en común, se necesitará una herramienta para procesar y guardar esa información con el fin de utilizarla y analizarla lo mejor posible.

Para ello se utiliza la herramienta ArangoDB, una novedosa base de datos no SQL, multi modal y open source, desarrollada por triAGENS GmbH, empresa alemana especializada en soluciones que demandan un rápido acceso a grandes cantidades de datos [18].

Características

Arango DB es una base de datos no SQL, es decir difiere del modelo clásico de Sistema de Gestión de Bases de Datos Relacionales en aspectos como que no se utiliza en lenguaje SQL para realizar las consultas y los datos que se almacenan en la misma no precisan ser una estructura fija determinada, como por ejemplo, tablas.

Además, ArangoDB es una base de datos que sirve documentos a clientes. Estos documentos son transportados usando el formato JSON y una conexión TCP y el protocolo HTTP.

Por otra parte, ArangoDB incluye una interfaz web, llamada Aardvark, la cual proporciona una interfaz gráfica al usuario para que su uso sea más sencillo. También cuenta con una Shell interactiva llamada Arangosh.

En cuanto al modelo de datos usado, podríamos realizar la siguiente comparativa para entenderlo mejor:

- Un documento correspondería a una fila en una base de datos relacional
- Una colección sería la tabla

La diferencia es que en un RDBMS tradicional se debe definir las columnas antes de que se puedan guardar datos en la tabla, como ArangoDB es "schema-less", significa que cada documento puede tener una estructura completamente diferente y estar guardado junto con otros documentos en la misma colección.

Documentos en ArangoDB

- ✓ Contienen cero o más atributos con un valor, que puede ser atómico, number, string, null o de tipo compuesto, array, documento, objeto.
- ✓ Un documento en ArangoDB es un objeto JSON
- ✓ Cada documento tiene una clave primaria única que lo identifica dentro de su colección
- ✓ Cada documento es identificado únicamente por de su "document handle" en todas las colecciones de una misma base de datos.
- ✓ No se necesita definir que atributos puede tener un documento

Ilustración 10 : Data model

Colecciones en ArangoDB

Los documentos se agrupan en colecciones, así una colección contiene cero o más documentos.

Hay dos tipos de colecciones: colecciones de documentos y colecciones Edge, estas últimas guardan documentos también pero incluyen dos atributos especiales _from y _to, que sirven para crear relaciones entre documentos.

Las colecciones existen dentro de las *bases de datos,* puede haber una o varias bases de datos y la base de datos por defecto es _system la cual no puede ser borrada.

AQL Arango Query Languague

El lenguaje para queries de Arango (AQL) permite recuperar y modificar datos que están guardados en las bases de datos de ArangoDB. En un lenguaje principalmente declarativo, es decir, que las queries expresan qué resultado quiere conseguirse, pero no cómo debe ser conseguido. Entre sus ventajas encontramos que es un lenguaje que busca que sea fácilmente legible y muchas palabras principales son adquiridas del inglés.

Para facilitar su uso, ArangoDB proporciona patrones comunes que se utilizan en las queires y adjunta ejemplos. Las palabras clave utilizadas son las siguientes:

-FOR: iteración en array

-RETURN: proyecta el resultado

-FILTER: para filtrar resultados

-SORT: para ordenar resultados

-LIMIT: cortar o limitar resultados

-LET: asignación de variable

-COLLECT: agrupar resultados

-INSERT: inserción de nuevos documentos

-UPDATE: actualización de documentos existentes

-REPLACE: reemplazo de documentos existentes

-REMOVE: borrado de documentos existentes

UPSERT: inserción o actualización de documentos existentes

```
FOR u IN users
  FILTER u.type == "newbie" && u.active == true
  RETURN u.name
```

Ilustración 11: Ejemplo query

4.1.3. Gephi

Este software libre se utiliza para la visualización de redes y grafos de todo tipo. Es una herramienta muy útil para analizar datos, ya que los muestra mediante gráficos y grafos llamativos e intuitivos, por otro lado, el usuario tiene la posibilidad de manipular e interactuar con las estructuras de los grafos para así poder estudiar y comprender los patrones que se describen.

Utiliza una serie de algoritmos de análisis y diseño para realizar los grafos, además de herramientas para interactuar con ellos. Entre las utilidades de Gephi destacan:

- Análisis exploratorio de datos
- Análisis de vínculos
- Análisis de redes sociales
- Análisis de redes biológicas

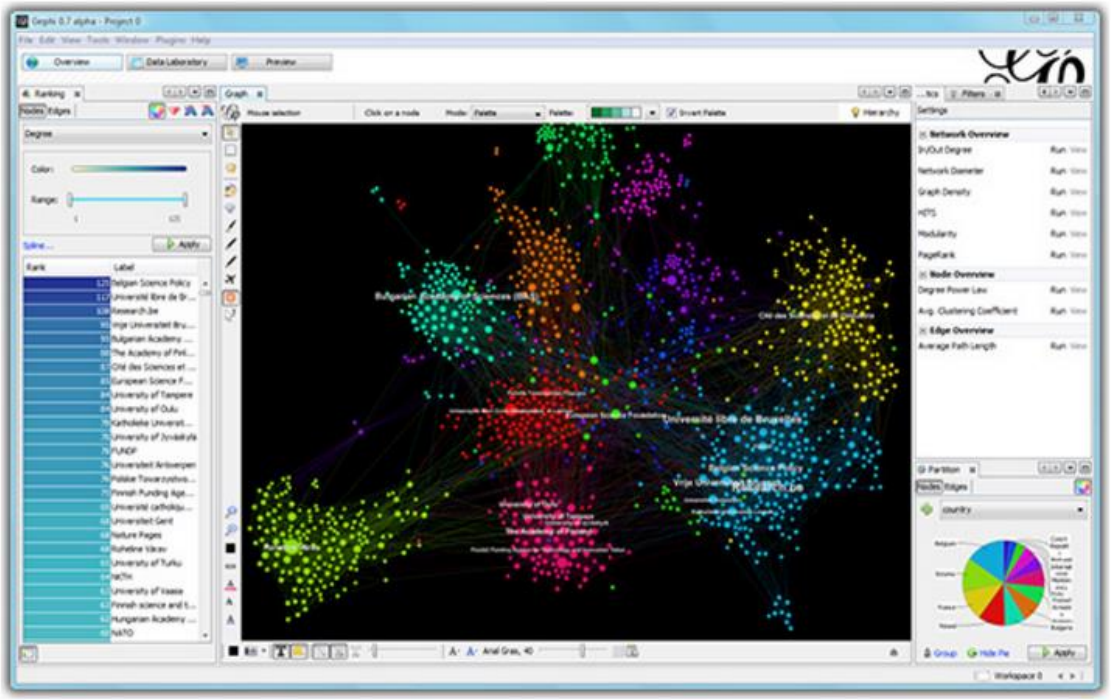


Ilustración 12: Gephi [19]

4.1.4. MonkeyLearn

Es una herramienta de aprendizaje automático, con plataforma en la nube, que sirve para analizar textos [20]. Permite tanto a usuarios como a compañías acceder fácilmente a datos de un texto "en bruto" y analizar dicho texto para detectar, por ejemplo, temas o el sentimiento expresado en tweets, chats, artículos…etc.

MonkeyLearn proporciona las siguientes utilidades:

- ∂ Una interfaz gráfica en la web que permite utilizar la herramienta de forma sencilla, pudiendo crear y probar algoritmos de aprendizaje automático para resolver problemas específicos o también se pueden utilizar algoritmos ya creados para casos más comunes como el análisis de sentimiento o detección de temas.
- ∂ Su plataforma de computación permite probar los algoritmos de aprendizaje automático al instante sin tener que instalar ningún software adicional.
- ∂ Con su API y SDK permite integrar MonkeyLearn con cualquier proyecto software usando casi cualquier lenguaje de programación.

Una de las características más importantes de esta herramienta es que podemos entrenar el algoritmo de aprendizaje automático con nuestros datos particulares. Esto nos va a ser de gran utilidad a la hora de mejorar la precisión para utilizar MonkeyLearn para clasificación de textos, detección de temas y análisis de sentimiento.
Las funcionalidades de MonkeyLearn se organizan en tres módulos:

- Clasificación: Módulo que toma el texto y devuelve etiquetas o categorías organizadas en una jerarquía.
- Extracción: Este módulo extrae datos particulares sin un texto, por ejemplo, entidades, direcciones o palabras clave.
- Pipelines: Este módulo es una combinación de los anteriores para construir procesos de alto nivel.

5. Diseño e implantación del sistema

En este apartado se explicarán los pasos que se han seguido para desarrollar nuestro trabajo. Tras analizar qué se quiere conseguir y cómo se va a lograrlo, es hora de poner en práctica los conocimientos que se tienen y desarrollar la herramienta propuesta.

Se comenzará con un esquema explicativo de la metodología que se ha seguido en la realización del proyecto, para posteriormente explicar cada paso realizado en el diseño del mismo.

Como se puede ver, el proyecto está estructurado en las siguientes fases:

- Extracción y almacenamiento de un corpus de tweets
- Creación de un motor de evaluación del sentimiento
- Análisis y extracción de conclusiones de los resultados obtenidos

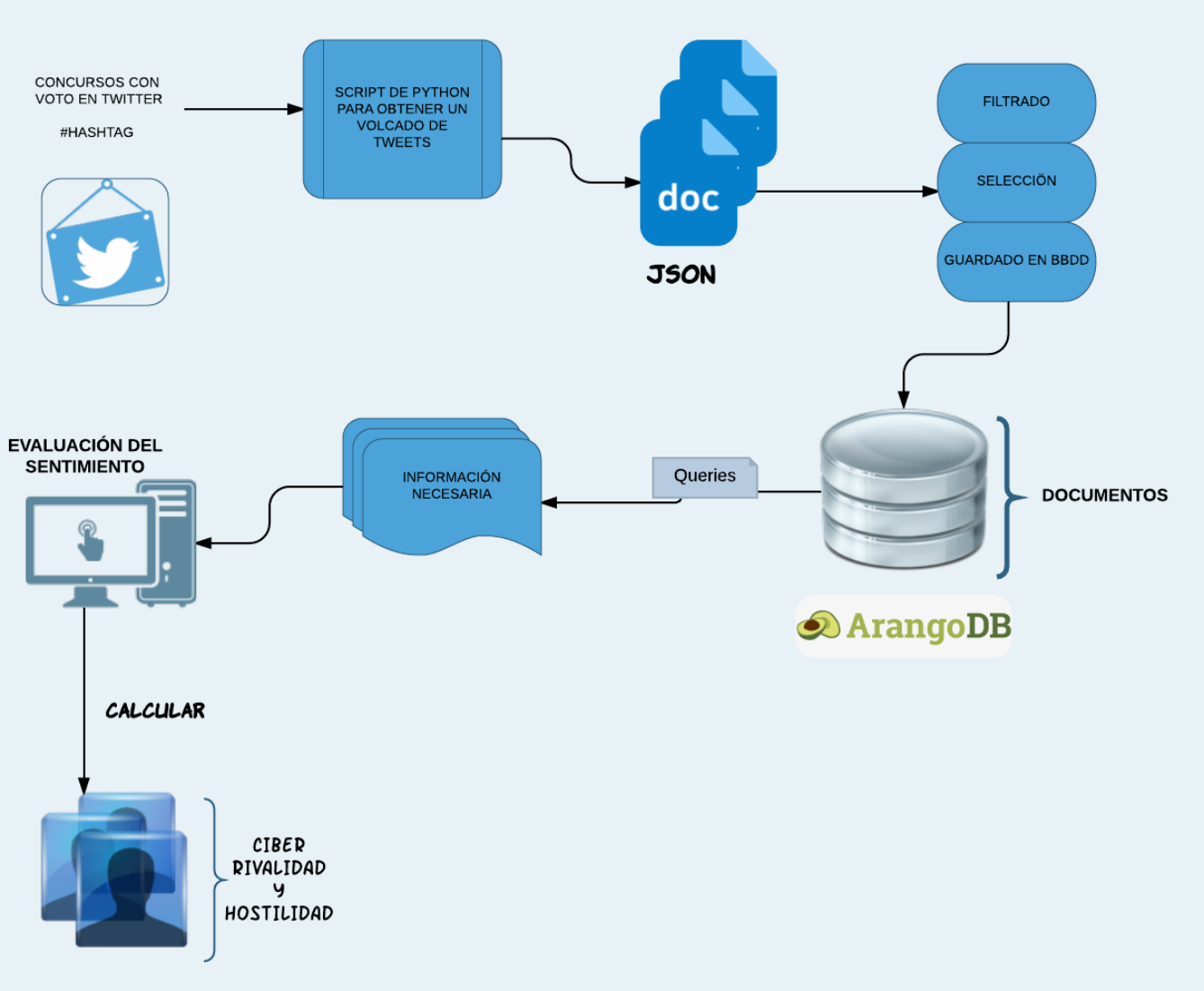
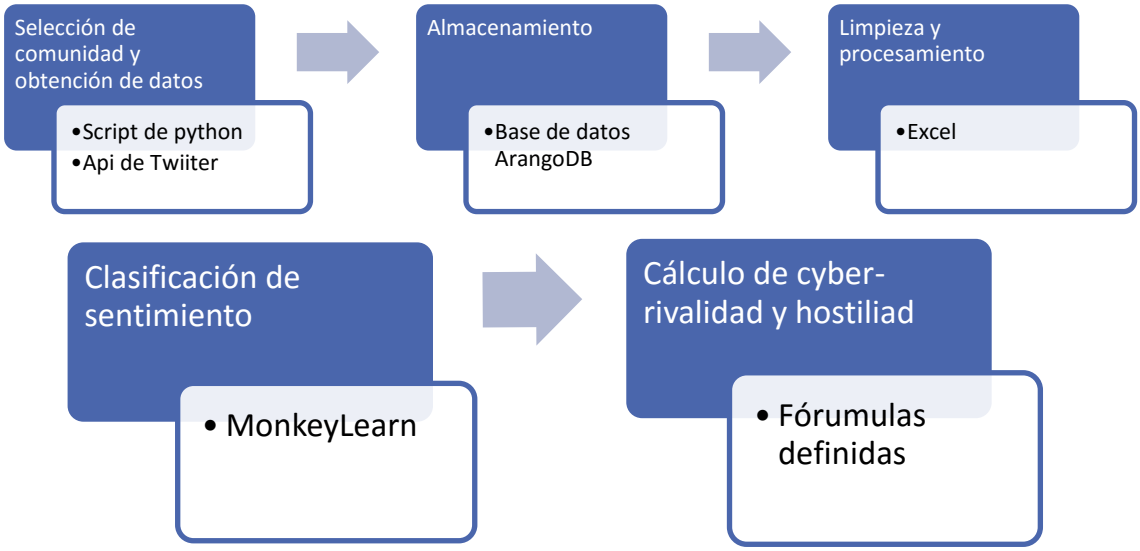


Ilustración 13: Diseño del sistema

Con el objetivo de clarificar la relación entre el análisis del sistema y su diseño, se realiza a continuación un esquema del mapeo del proceso tecnológico seguido:



5.1. Obtención de datos

Como se explicaba al comienzo del presente trabajo, el análisis que se va a realizar será sobre la red social Twitter, obteniendo así un corpus o conjunto de tweets determinados para su posterior análisis.

Este es el primer paso en el desarrollo de nuestro sistema y para ello nos hemos apoyado en las herramientas descritas anteriormente, API de Twitter (Apartado 3.1.2) y un script en lenguaje de programación Python (Apartado 4.1.1) .

A la hora de conseguir los datos, se ha elegido Twitter y su API porque permite acceder a una gran cantidad de información de carácter público como son los tweets generados por los usuarios, sin embargo se han encontrado con ciertas limitaciones que en principio no han supuesto grandes inconvenientes a la hora de desarrollar el proyecto,como son que a la hora de recuperar los tweets estos deben hacer referencia a un hashtag del que se esté twiteando actualmente, ya que su recuperación será en tiempo real, por lo que si se quieren tweets de algún tema del pasado o del que no haya muchos usuarios hablando, el script producirá un error ya que no encontrará datos.

Para utilizar la API de Twitter debemos darnos de alta como desarrolladores y crear una aplicación, esto se hace de manera sencilla ya que sólo necesitamos tener una cuenta de Twitter creada la cuál utilizaremos para obtener una serie de parámetros que nos piden para utilizar la API.

Una vez obtenidas estas claves, las añadiremos en el fichero de configuración de nuestro programa Python para que este se comunique con la API.

En cuanto al funcionamiento del script utilizado, este obtendrá una lista de tweets que serán extraídos en formato JSON a una carpeta de destino especificada. Como parámetros a introducir tendremos el hashtag que contendrán todos los tweets extraídos y el número de tweets que queremos recuperar.

```
>twitter_stream_download.py -q PremiosMTVMIAW -d salida -n 100
```

Como salida obtendremos un archivo JSON con el conjunto de tweets recuperados para ese hashtag concreto.

Los tweets recuperados contienen todos los atributos propios de un tweet, siguiendo la estructura de almacenamiento JSON, como explicamos en el análisis del sistema (Apartado 3.1.3) , no todos los campos nos van a ser útiles, por lo tanto más adelante seleccionaremos aquellos que utilizaremos para extraer información.

5.2. Almacenamiento y procesamiento de datos

Una vez obtenidos los datos que se van a utilizar para el análisis, los almacenaremos creando una base de datos utilizando la herramienta Arango DB , la cual permite crear colecciones de documentos.

Por lo tanto, se decide crear colecciones que almacenen todos los tweets que se han recuperado de un hashtag concreto, para crearlas se utiliza la interfaz gráfica que proporciona Arango DB y la Shell.

El primer paso tras instalar y configurar Arango DB será la creación de una nueva base de datos que contendrá las colecciones y los documentos. Para crearla basta con pulsar el botón de add database e introducir el nombre que le pondremos a la misma y el usuario que la crea, que por defecto será root.

Lo siguiente que se necesita será una colección en la que insertar documentos, también la crearemos mediante la interfaz web de Arango, en la pestaña de collections pulsando add collection y añadiendo su nombre.

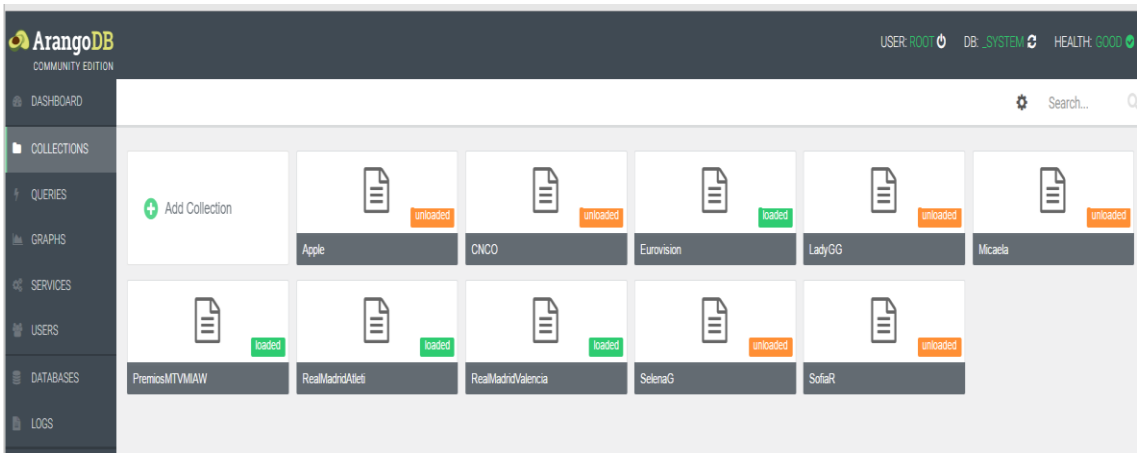


Ilustración 14: Ejemplo interfaz gráfica ArangoDB

A continuación se añaden los documentos a nuestra colección, para ello se utilizará la Shell de Arango DB ya que se necesita utilizar un comando que permite importar la gran cantidad de datos en el archivo JSON. En la siguiente ilustración podemos ver el funcionamiento del comando arangoimp, el parámetro file indica el nombre del archivo a importar, type el tipo de formato del mismo, nosotros utilizaremos archivos JSON, y por último el nombre de la colección en la que queremos insertar los documentos.

Una vez ejecutado se muestra una breve estadística para verificar que todo ha salido correctamente, y en la que podemos ver los documentos creados.

```
C:\Program Files\ArangoDB3 3.1.18\usr\bin>arangoimp --file "mtv.json" --type json --collection MTV
Please specify a password:
Connected to ArangoDB 'http+tcp://127.0.0.1:8529', version 3.1.18, database: '_system', username: 'root'
----------------------------------------
database:               _system
collection:             MTV
create:                 no
source filename:        mtv.json
file type:              json
connect timeout:        5
request timeout:        1200
----------------------------------------
Starting JSON import...
2017-04-28T17:30:40Z [10260] INFO processed 22293 bytes (3%) of input file

created:          4
warnings/errors:  0
updated/replaced: 0
ignored:          0
```

Ilustración 15: Ejecución de arangoimp

Hecho esto, se tendrá una colección compuesta por documentos, cada documento es un tweet que sigue la estructura de almacenamiento JSON y que guarda todos los atributos que ha proporcionado la API de Twitter y de los que se ha hablado en el apartado 3.1.2.

En las siguientes imágenes se muestra un ejemplo de cómo aparece una colección en la interfaz gráfica y el aspecto que tiene un documento guardado en la misma y con el que más adelante se trabajará.

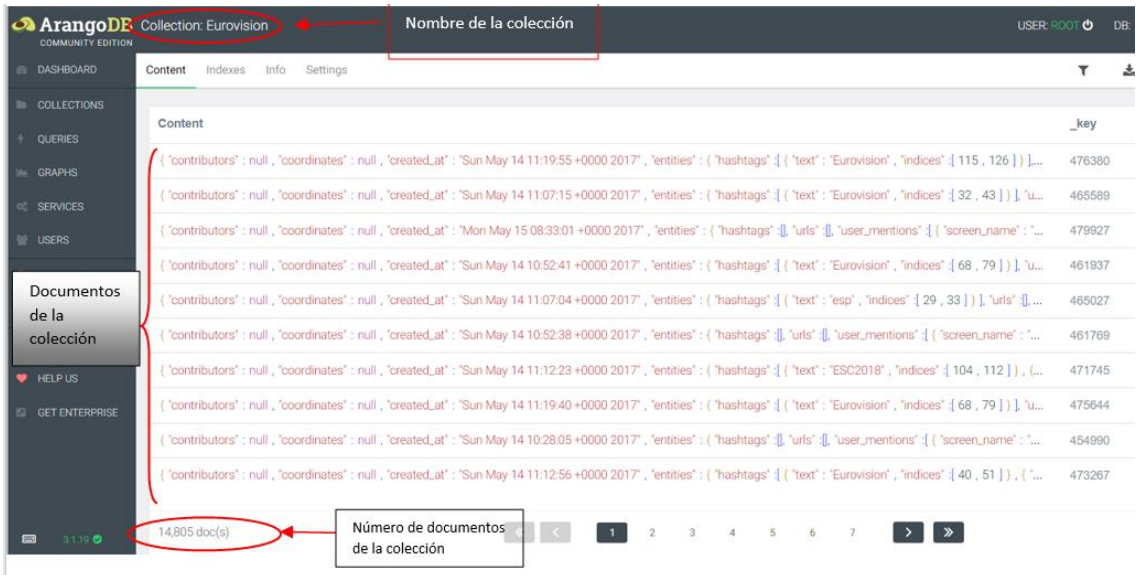


Ilustración 16: Ejemplo de colección en ArangoDB

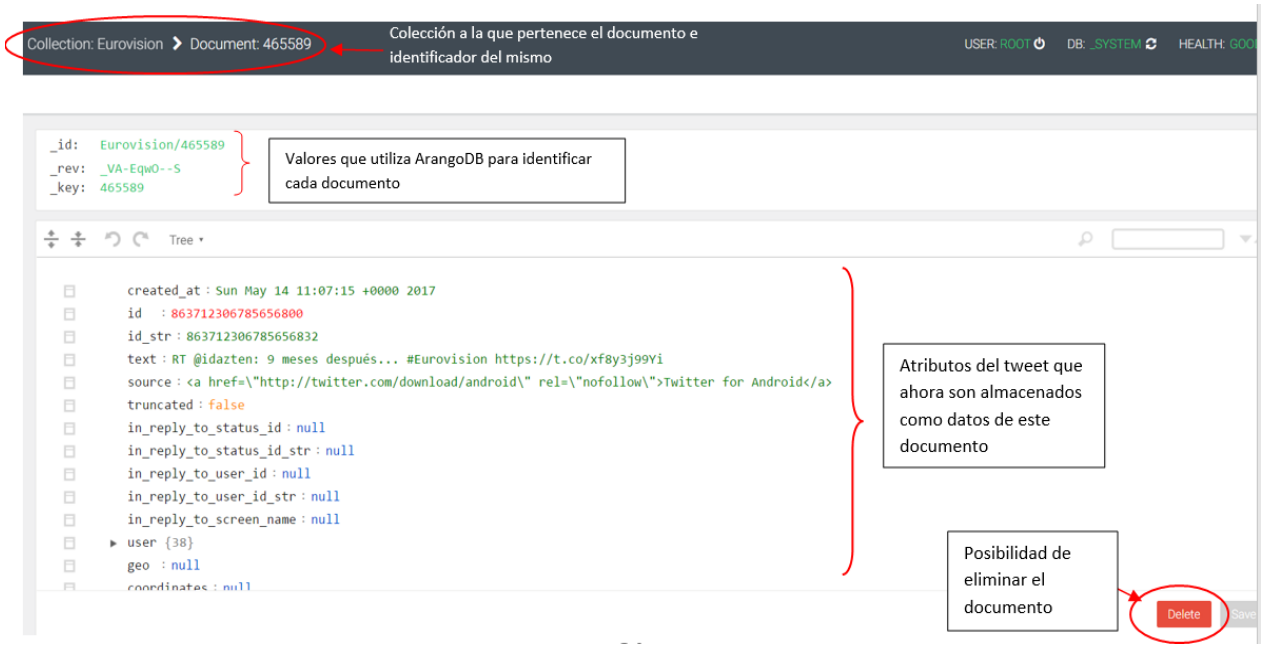


Ilustración 17: Ejemplo de documento en ArangoDB

5.3. Creación de un motor de evaluación del sentimiento

Una vez se tienen los datos almacenados en nuestra base de ArangoDB es hora de tratarlos para obtener su polaridad. Para conseguir esto se acudirá a una herramienta externa, MonkeyLearn, que es capaz de crear un modelo de aprendizaje con el que analizaremos los tweets. A continuación, se explicará, los pasos seguidos en la utilización de esta herramienta y su funcionamiento.

[1]. Creación de un módulo

Lo primero que se debe hacer es crear un módulo, que sería como un proyecto, se definirá su nombre, descripción, si se quiere que sea público o privado y el tipo de módulo, en nuestro caso vamos a crear un clasificador.

Se definirá también el tipo de problema que se quiere resolver, ya que esto condicionará de cara a obtener mejores resultados, nuestra opción será “Social Media”.

A continuación, se definirá el lenguaje que se va a usar, inglés, y el tipo de textos que se van a clasificar, tweets.

[2]. Creación de un category Tree (árbol de categorías)

Con un árbol de categorías se podrán organizar las etiquetas que se quieren asignar a nuestros textos de manera jerárquica. Para nuestro módulo de análisis de sentimiento utilizaremos las categorías positivo y negativo.

Se necesitarán ejemplos de cada categoría para que MonkeyLearn pueda crear un modelo fiable, para ello se utilizarán un conjunto de tweets ya clasificados como

negativos o positivos en formato CSV y gracias al GUI de la herramienta se podrán cargar en nuestro árbol de manera sencilla. En la siguiente ilustración se muestra un ejemplo en el que podemos visualizar cómo el interfaz nos da la opción de que parte del fichero vamos a utilizar como texto y cual como categoría, de esta forma señalaremos el contenido del tweet como el texto y la etiqueta positivo o negativo será la categoría.

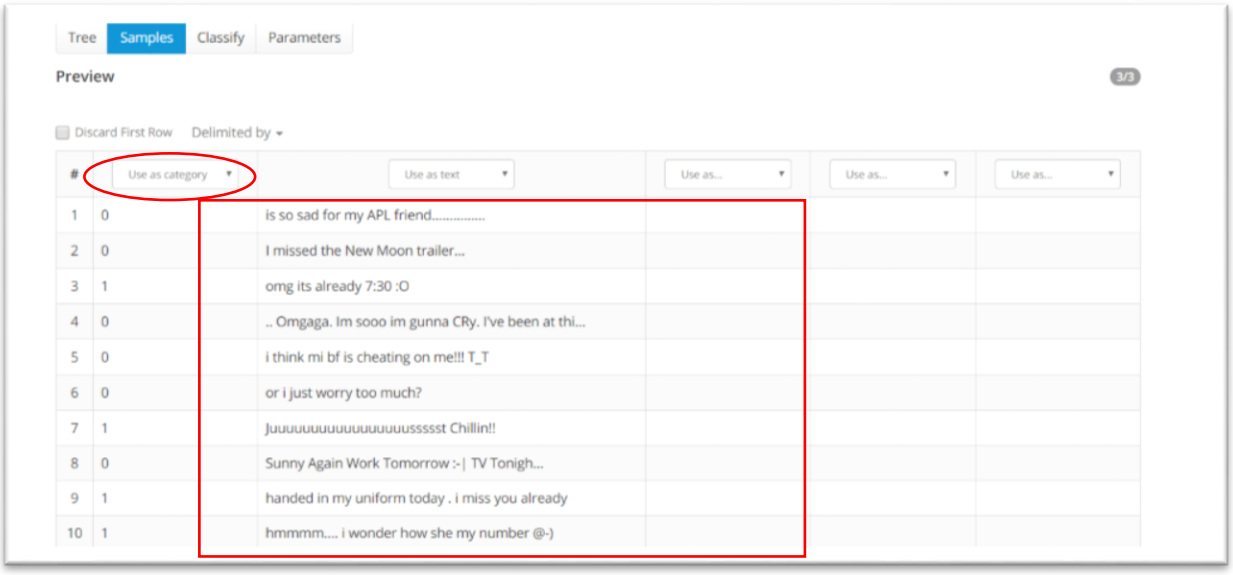


Ilustración 18: Carga de datos

[3]. Ajuste de parámetros de clasificación

Estos parámetros servirán para customizar más nuestro modelo de cara a conseguir mayor precisión. En nuestro trabajo se utilizarán los que vienen definidos por defecto, utilizando como idioma el inglés y añadiendo algunas "stopWords" que son aquellas que creemos que no aportan nada a la clasificación del texto (artículos, conectores).

En cuanto al algoritmo de clasificación, se usará Multinomial Naive Bayes, [21], el cual estima la probabilidad de una palabra dada en una clase C, como la frecuencia relativa del término t en los documentos que componen la clase. La variación tiene en cuenta el número de ocurrencias del término t en los documentos de entrenamiento de la clase C, incluyendo ocurrencias múltiples.

[4]. Ejecución del modelo de entrenamiento

Una vez se tiene el modelo creado se debe “entrenarlo” con los ejemplos cargados para que sea capaz de clasificar los textos que le introduzcamos. En la pestaña etiquetada como “sandbox” se pulsa el botón “train” y comenzará a funcionar. Una vez terminado el proceso se mostrará la siguiente imagen:

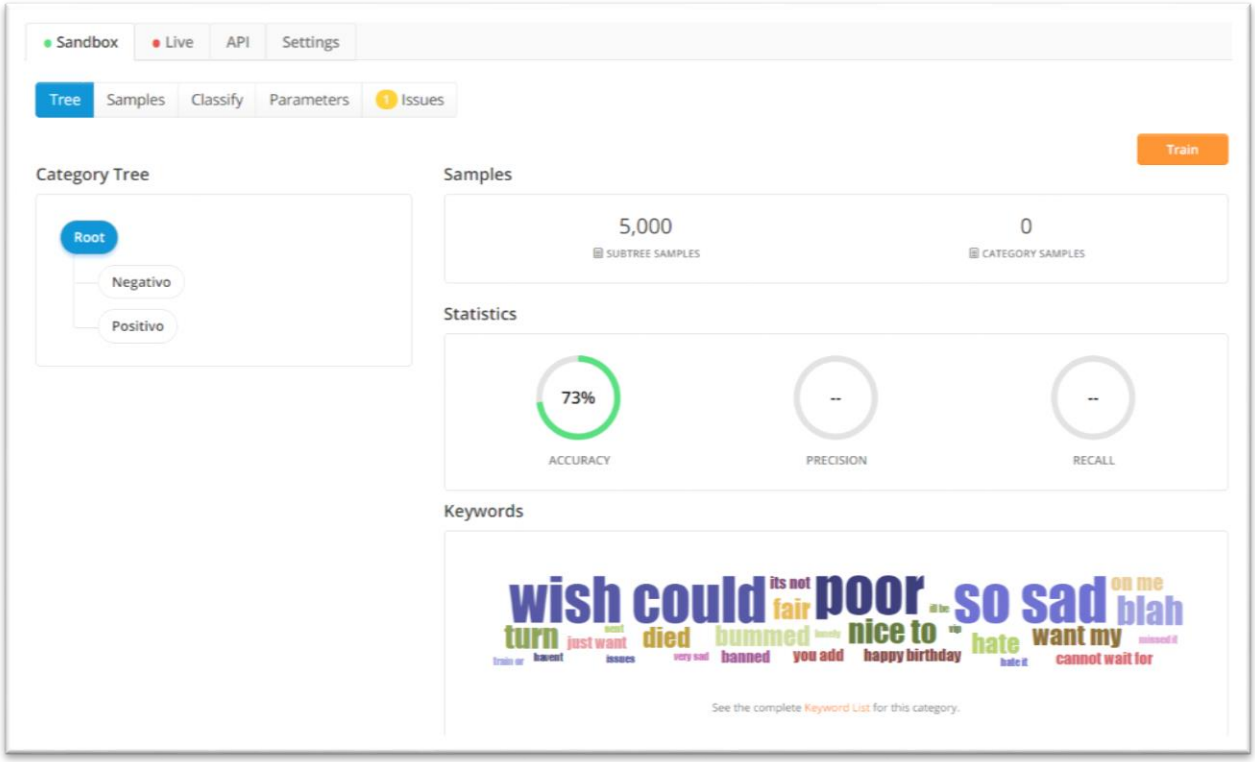


Ilustración 19: Modelo de entrenamiento

El modelo se compone de distintas pestañas, tree,samples,classify,parameters y la última issues aparece como notificación para realizar posibles mejoras al modelo.

Se explicará, a continuación, con detalle los apartados que componen el modelo para así entender mejor como funciona:

Samples

Contabiliza los ejemplos utilizados para el entrenamiento del modelo, como se utiliza una versión gratuita está limitado a 5000 ejemplos. Si se pulsa en una categoría concreta aparecerán los ejemplos de cada una con sus estadísticas, de las que se hablará más adelante. Además, se puede ver cómo han sido clasificados los distintos ejemplos que se han utilizado, el porcentaje de positividad y negatividad con el que cuenta ese ejemplo, la etiqueta que se le ha asignado y las palabras con influencia positiva o negativa según el clasificador.

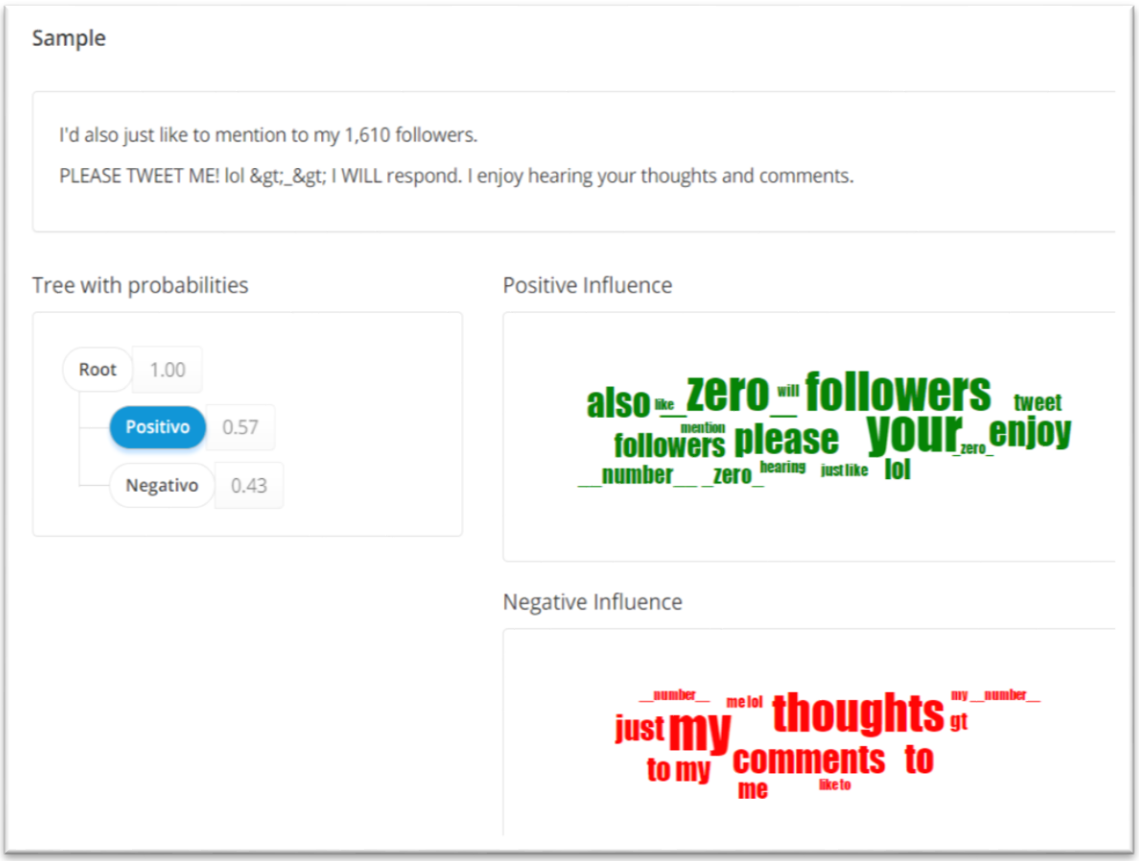


Ilustración 20: Sample

Statistics

Este apartado es uno de los más interesantes y que más información nos aporta, por lo que se detallará con más extensión. Las medidas que se encontrán son:

-Accuracy (Exactitud): Representa el porcentaje de ejemplos que han sido predichos en la categoría correcta. Esta métrica muestra cómo de buena es la distinción que una categoría "padre" distingue entre sus categorías "hijos". En nuestro caso es simple puesto que nuestro árbol solo tiene dos categorías hijo y estas no contienen subcategorías a la vez.

- Precision y Recall: La precisión para una categoría no raíz representa el porcentaje de ejemplos en el test que han sido clasificados en esta categoría por su padre y que realmente pertenecen a la misma. Este valor solo tiene sentido en etiquetas que no son raíz

En cuanto al Recall, es el porcentaje de todos los ejemplos que originalmente pertenecían a esa categoría y que en el proceso de evaluación fueron clasificados correctamente en esa categoría por su nodo padre. Como en el recall, solo se da en nodos no raíz.

Ambas medidas son útiles para comprobar la exactitud de cada categoría hoja o hijo.

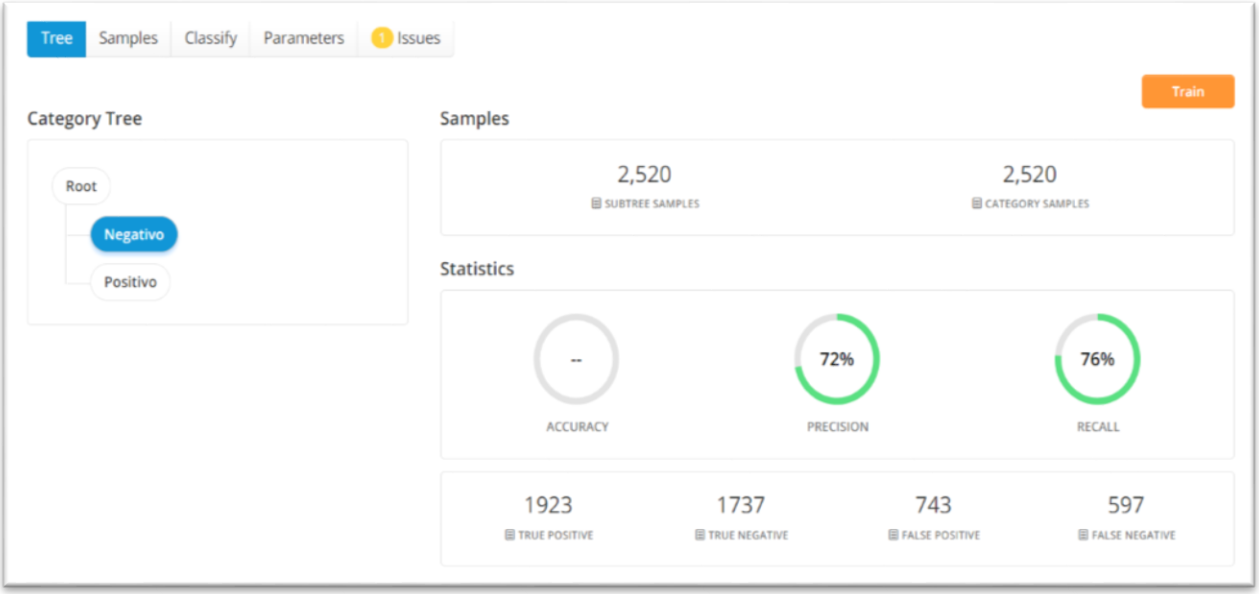


Ilustración 21: Precisión y Recall

Como se puede observar en la ilustración, se muestra el número de true /false positives/negatives. Estos datos hacen referencia a la precisión y recall, el número de true positives indica el número de ejemplos que han sido identificados correctamente como positivos, mientras que el false positive indica lo contrario, aquellos ejemplos que fueron clasificados como positivos, pero no lo son.

Ocurre lo mismo para los ejemplos clasificados como true y false negative, sólo que estos fueron clasificados como negativos.

Confusion Matrix

La matriz de confusión muestra la confusión entre la categoría actual donde el ejemplo está categorizado y a la que pertenece, y la categoría predicha, es decir, la categoría que predijo el clasificador. En nuestro caso tenemos 597 ejemplos que fueron clasificados como negativos pero fueron predichos como positivos. Los números que aparecen en la diagonal central de la matriz deben ser siempre mayores que 0 lo que indicará que nuestro modelo funciona con buena precisión. Los números rojos que son altos indican errores y son buenos candidatos para mejorar nuestro modelo.

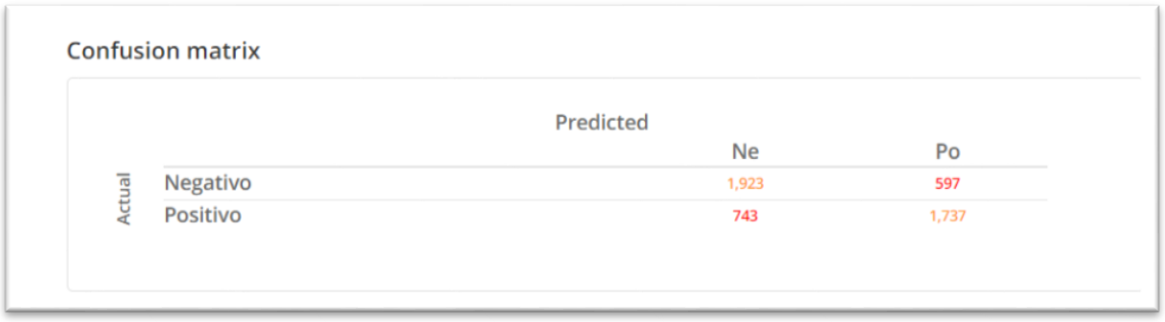


Ilustración 22: Confusion matrix

Además, MonkeyLearn incorpora una opción de mejora de esta matriz, y es que da la opción de clasificar nosotros mismos los ejemplos que se consideran confusos o mal clasificados, y por lo tanto no permite colocarlos en la categoría que creemos más apropiada.

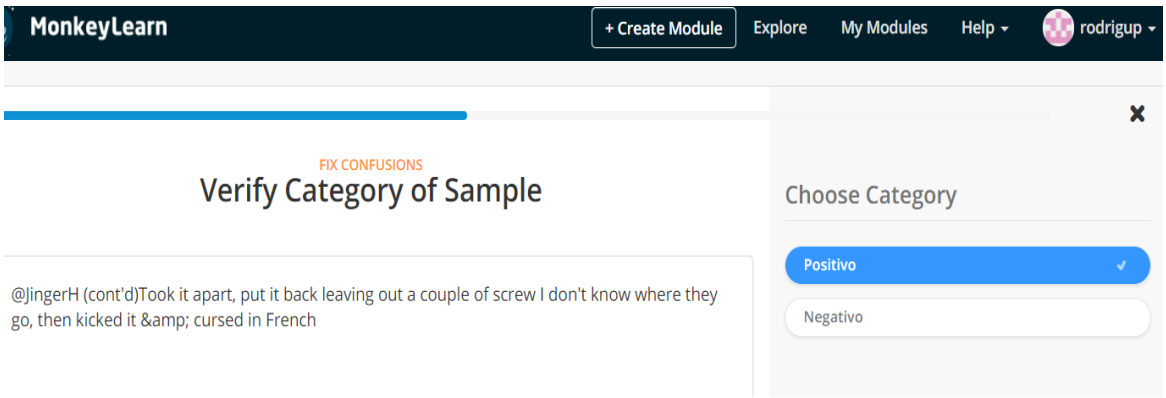


Ilustración 23: Solventar confusiones en MonkeyLearn

[5]. Clasificación de los tweets almacenados

Una vez se tiene nuestro modelo creado y configurado se va a utilizar con el conjunto de tweets de las distintas comunidades que se han almacenado en ArangoDB. Para extraer los datos necesarios se hará uso de AQL ya que no todos los datos almacenados en un tweet van a ser de utilidad. A continuación se muestra la query que se ha realizado para obtener el usuario que escribió el tweet, el texto que se quiere que clasifique y los hashtags y menciones que realizó, además de escoger sólo los tweets que estén en inglés.

Ilustración 24: Query

La sintaxis que se describe en la query siginifica que, para todos los documentos en una colección, aquí aparece el nombre de una de ellas, se filtrarán los resultados por el parámetro lenguaje que será inglés, y se devolverá el nombre del usuario, el texto del tweet, en el array hashtag se devolverán todas las posiciones del array [*] y en concreto el campo texto, con las menciones se hará lo mismo pero se obtendrá el campo nombre.

Una vez realizada la query se obtendrá un archivo en formato JSON que se convertirá a Excel para utilizarlo en la herramienta MonkeyLearn.

A continuación, en el apartado Classify de nuestro módulo, se cargará el fichero a analizar y como salida resultará un archivo Excel o csv con todos los datos que tenía anteriormente el documento más un nuevo dato que será polaridad y un tanto por ciento.

En la imagen se muestra el aspecto que tiene el clasificador de textos, en el cual debemos seleccionar el campo que queremos utilizar como texto a clasificar y que será aquel que contiene el texto del tweet. Los demás campos se mantendrán cuando el clasificador cree el fichero de salida, pero para clasificar los desecharemos puesto que nuestro objetivo es obtener la polaridad del texto del tweet.

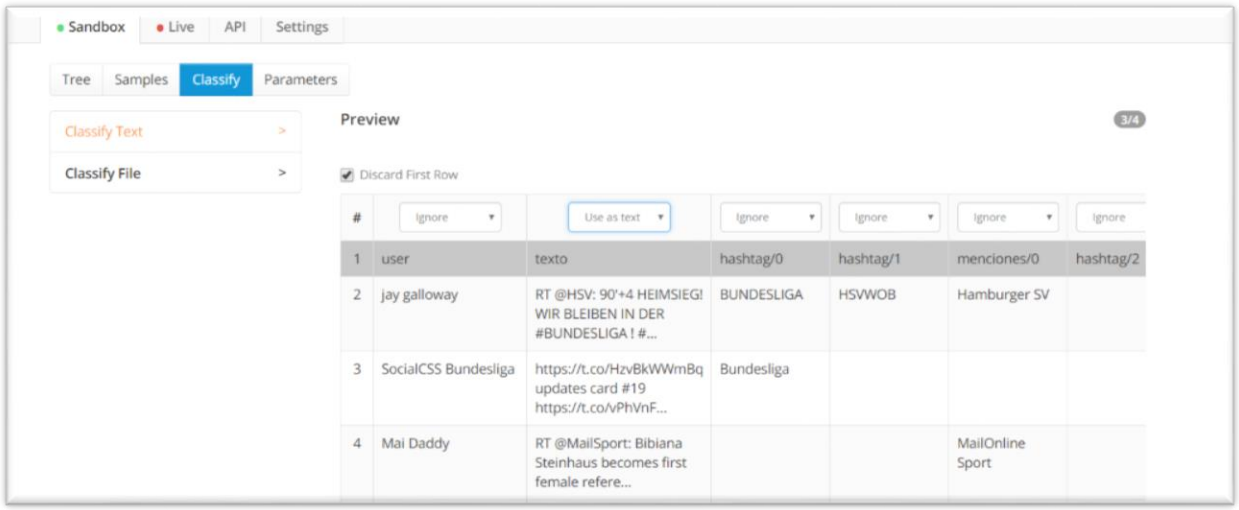
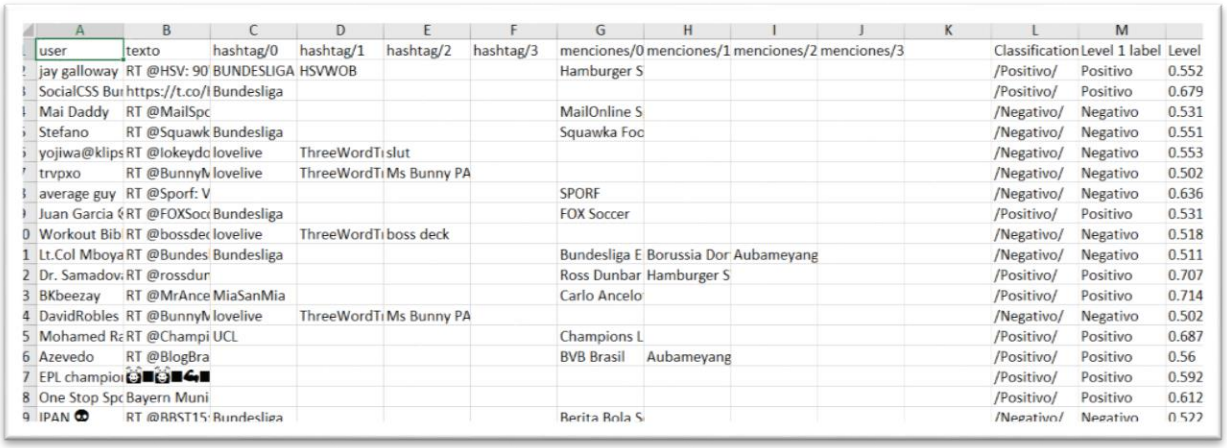


Ilustración 25: Clasificador

A continuación, el clasificador se pondrá en funcionamiento generando un archivo Excel, en él aparece el usuario que ha posteado el tweet, el texto que escribió, los hashtags y menciones y su clasificación como positivo o negativo, su etiqueta correspondiente y el porcentaje de probabilidad de pertenecer a esa categoría:



	A	B	C	D	E	F	G	H	I	J	K	L	M	
	user	texto	hashtag/0	hashtag/1	hashtag/2	hashtag/3	menciones/0	menciones/1	menciones/2	menciones/3		Classification	Level 1 label	Level
2	jay galloway	RT @HSV: 90	BUNDESLIGA	HSVWOB			Hamburger S					/Positivo/	Positivo	0.552
3	SocialCSS Bu	https://t.co/I	Bundesliga									/Positivo/	Positivo	0.679
4	Mai Daddy	RT @MailSpc					MailOnline S					/Negativo/	Negativo	0.531
5	Stefano	RT @Squawk	Bundesliga				Squawka Foc					/Negativo/	Negativo	0.551
6	yojiwa@klips	RT @lokeydo	lovelive	ThreeWordTi	slut							/Negativo/	Negativo	0.553
7	trvpxo	RT @BunnyM	lovelive	ThreeWordTi	Ms Bunny PA							/Negativo/	Negativo	0.502
8	average guy	RT @Sporf: V					SPORF					/Negativo/	Negativo	0.636
9	Juan Garcia (	RT @FOXSoc	Bundesliga				FOX Soccer					/Positivo/	Positivo	0.531
10	Workout Bib	RT @bossdec	lovelive	ThreeWordTi	boss deck							/Negativo/	Negativo	0.518
11	Lt.Col Mboya	RT @Bundes	Bundesliga				Bundesliga E	Borussia Dor	Aubameyang			/Negativo/	Negativo	0.511
12	Dr. Samadov	RT @rossdur					Ross Dunbar	Hamburger S				/Positivo/	Positivo	0.707
13	BKbeezay	RT @MrAnce	MiaSanMia				Carlo Ancelo					/Positivo/	Positivo	0.714
14	DavidRobles	RT @BunnyM	lovelive	ThreeWordTi	Ms Bunny PA							/Negativo/	Negativo	0.502
15	Mohamed Ra	RT @Champi	UCL				Champions L					/Positivo/	Positivo	0.687
16	Azevedo	RT @BlogBra					BVB Brasil	Aubameyang				/Positivo/	Positivo	0.56
17	EPL champio											/Positivo/	Positivo	0.592
18	One Stop Spo	Bayern Muni										/Positivo/	Positivo	0.612
19	IPAN	RT @BBST15	Bundesliga				Berita Bola S					/Negativo/	Negativo	0.522

Ilustración 26: Excel resultado

5.3.1. Componente alternativo

En la realización del presente proyecto se ha tenido en cuenta la complejidad y delicadeza del tema del análisis del sentimiento. En relación a ello, se propone una alternativa al clasificador utilizado, pero que finalmente no se utilizó para el proyecto.

Este componente se desarrolló en java, y contaba con la ventaja de que el corpus de entrenamiento era mucho mayor que el que la herramienta MonkeyLearn permitió utilizar, ya que se contaba con la limitación de la versión gratuita. Dicho componente trabajaba con un conjunto clasificado de un millón de tweets, cuyas etiquetas correspondían a 0 para los tweets negativos y 1 para los positivos.

El sistema permitía clasificar tanto un texto introducido por parámetro, como un conjunto de textos, previa modificación del código java del mismo.

Como salida, se muestra el tweet o texto introducido como positivo o negativo.

Se presenta este clasificador a modo de alternativa, sin embargo, se llegó a la conclusión de que no era del todo fiable y preciso, debido a que la salida indicaba únicamente las etiquetas “positivo” y “negativo” y para poder utilizar esa clasificación de manera eficiente en nuestro proyecto, se debía aplicar un rango porcentual aleatorio, debido a que el componente no mostraba un tanto por ciento de positividad o negatividad como sí que muestra la herramienta utilizada, MonkeyLearn.

Es por este motivo por el cual se decidió no utilizar este sistema para clasificar los tweets, puesto que, aunque como ventajas contaba con un modelo de entrenamiento muy numeroso, como inconvenientes tenía la imprecisa clasificación del sentimiento, motivo que se determinó como de suficiente valor para no utilizarlo.

5.4. Calculo de cyber-rivalidad y hostilidad

Este es uno de los pasos más importantes para nuestro estudio, ya que permitirá obtener resultados numéricos que serán analizados y mediante los cuales se podrán sacar conclusiones.

Una vez se tienen los datos en una tabla de Excel con los campos que se consideran útiles, se deben realizar algunas labores de enriquecimiento y limpieza de los mismos.

Lo primero que se hizo fue eliminar los campos que no son de utilidad y que están almacenados en la base de datos, para que en nuestra tabla Excel solo apareciera lo primordial.

Por otra parte, se realiza una limpieza de algunos tweets que contenían palabras en alfabetos que no fueran latinos, puesto que, aunque se filtraron los datos por idioma, en algunos casos aparecieron caracteres de otros alfabetos.

Para la definición de la comunidad a estudiar se explicó que se haría uso de un hashtag común que apareciera en todos los tweets posteados por los diferentes usuarios, sin embargo, también se necesita disgregar la propia comunidad para encontrar dos grupos que puedan ser potencialmente rivales. Esta tarea ha presentado ciertas limitaciones y dificultades, ya que es complicado encontrar un modo de saber si un usuario se está refiriendo expresamente a otro de una comunidad opuesta.

Por lo tanto, se tomó la decisión de plantear varios casos para definir comunidades dentro de una más global y así poder medir su rivalidad u hostilidad.

- ✓ Caso 1: Aparecen dos comunidades distintas en un mismo tweet.
  Ocurriría cuando se encontrarán usuarios que twittearan un texto en el que aparecieran menciones o hashtags que hicieran referencia a dos comunidades. Con esto podríamos deducir que el usuario es seguidor de una de las dos comunidades que menciona, y referencia a una segunda, con lo que podremos tratar de medir si existe rivalidad entre ambas comunidades mencionadas.
- ✓ Caso 2: Comunidad global y mención de subcomunidad. Para este caso, encontraríamos con que los usuarios mencionan a otra comunidad en sus tweets distinta de la global.
- ✓ Caso 3: Comunidad global.
  Se escogerá toda la comunidad diferenciada por su hashtag para realizar el cálculo de la hostilidad. Se aplicará esta opción para agrupar todas las subcomunidades dentro de una más grande.

Finalmente, para el cálculo de los valores finales que resultan de aplicar las fórmulas descritas en el apartado 3.3. Obtención de una fórmula , nos apoyamos de los datos proporcionados por la herramienta MonkeyLearn en la clasificación del sentimiento de los tweets. Según la ecuación 1 del citado apartado, la función $S(r_{C_i}^{i}, F_{C_k}^{descriptors})$ atiende a un valor que mide la polaridad de un tweet, para que los resultados tengan sentido hemos decidido obtener el valor de la función S de la siguiente manera :

- Etiqueta/Clasificación POSITIVA: Para los tweets determinados como positivos aplicaremos su porcentaje de positividad resultante en la evaluación de sentimiento, de manera que, para un tweet positivo, su valor de S será igual a su X % de positividad.
- Etiqueta/Clasificación NEGATIVA: Para los tweets clasificados como negativos su peso en la función S será de 0, debido a que será la forma de dar más valor a los tweets negativos en el cálculo de la rivalidad.

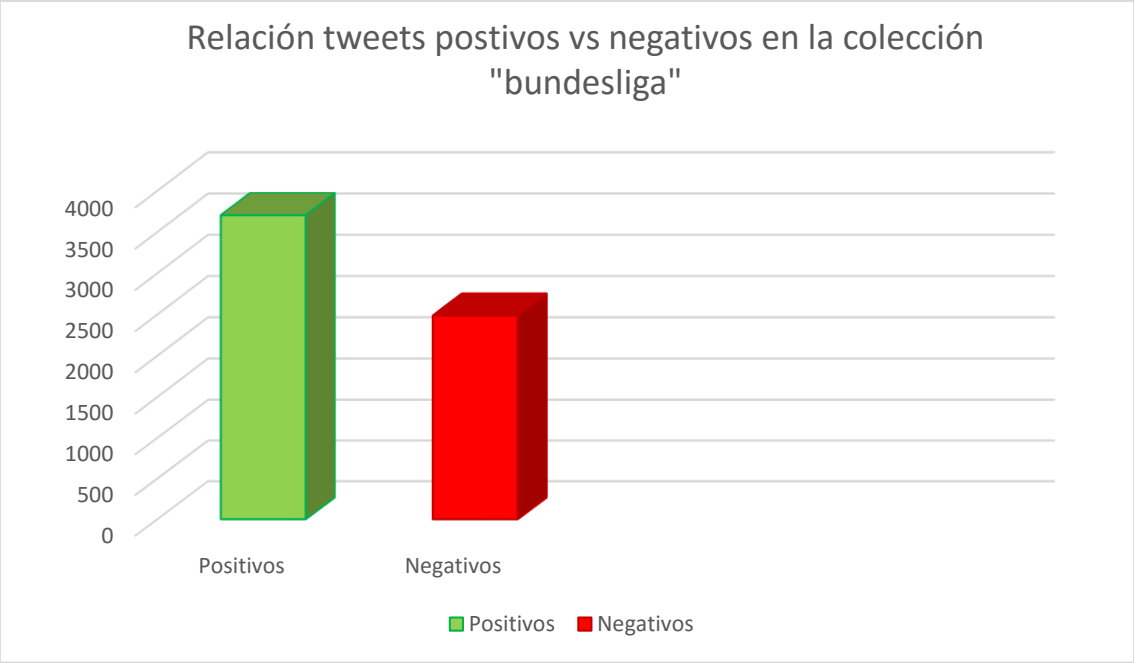


Ilustración 27: Tweets positivos vs negativos

6. Experimentación y Pruebas

Para verificar que nuestro sistema funciona y produce los resultados esperados, se han realizado dos acciones, la experimentación con los datos obtenidos con el fin de obtener las medidas buscadas, y las pruebas a nuestro sistema para confirmar su correcto funcionamiento.

6.1. Aplicación de fórmulas

Trabajaremos con varias colecciones de datos, de las que se ha hablado anteriormente en el apartado 3.2 Selección de la comunidad a estudiar, que hacen referencia a un hashtag concreto, y se estudiarán distintos casos.

Para cada hashtag organizaremos el experimento de la siguiente manera:

- Datos utilizados y filtrados:
  Se añadirá una breve explicación de cómo vamos a tratar el conjunto de datos que se obtuvieron anteriormente de acuerdo con las limitaciones que presenta cada comunidad, para así poder obtener resultados coherentes.

- Fórmulas aplicadas y resultados
  Se utilizarán las fórmulas descritas para obtener un valor numérico. Adicionalmente, se añadirán gráficas explicativas en los casos que resulte necesario.

Hashtag FACupFinal

Para esta comunidad correspondiente a la final de la liga inglesa, que enfrentaba a los equipos Chelsea y Arsenal, se ha aplicado el caso 1, obteniendo así descriptores que hacen referencia a ambos equipos en un mismo tweet. Para ello se han filtrado aquellos tweets que contengan un hashtag con las palabras “Arsenal”, “Chelsea”, o abreviaturas que les correspondan como AFC o CFC.

Datos

Una vez filtrados los datos, nuestro conjunto a analizar se reducirá sustancialmente de manera que tendremos unas 240 ocurrencias que corresponden a estas características.

user	texto	hashtag/0	hashtag/1	hashtag/2
Faseyi	Moses want to Dab 😂😂😂😂 #FACupFinal #Arsenal #Chelsea	FACupFinal	Arsenal	Chelsea
Nicholas Paliobagis	RT @Lil_Gilbino81: Arsenal are playing like Chelsea and Chelsea are pl...	FACupFinal	arsenal	chelsea
Âmır El.	RT @trtworld: #FACupFinal: #Arsenal have beaten #Chelsea 2-1 to clai...	FACupFinal	Arsenal	Chelsea
Ramin Leylabi	Could be the game changer unless arsenal get a second. #FACupFinal #...	FACupFinal	Arsenal	Chelsea
Portion Of Sport	@MarkFergYT & @JackSimmons101 are LIVE WATCHING THE #F...	FACupFinal	Arsenal	Chelsea
scott phillips®	Chelsea attacking the Arsenal defence #FACupFinal #Arsenal #Chelsea ...	FACupFinal	Arsenal	Chelsea
PKA GBCA	Corbyn 2 Russians 1. #FACupFinal #arsenal #chelsea	FACupFinal	arsenal	chelsea
Olayemi Agbeleye	RT @dwishank: Arsene Wenger every season 😂😂😂🤦#AFCvCFC #ars...	AFCvCFC	arsenal	chelsea
LALATE	GB⚽ #LALATE ⚽GB LIVE! #Arsenal vs #Chelsea 1-0 @ 63'! #FACup Winner #FACUPFINAL https://t.co/uQnBiFacJ0 https://t.co/yBlctvjdKP	LALATE	Arsenal	Chelsea
Dwishank	Arsene Wenger every season 😂😂😂🤦#AFCvCFC #arsenal #chelsea #...	AFCvCFC	arsenal	chelsea
Marc Williams	Thoroughly enjoyable game of football. Arsenal outstanding. #FACupFi...	FACupFinal	Arsenal	Chelsea
MarkFerg	RT @PortionOfSport_: @MarkFergYT & @JackSimmons101 are LI...	FACupFinal	Arsenal	Chelsea
Chris Walden	Haha FUCK OFF MOSES!! nice to see someone get punished for diving!...	FACupFinal	Arsenal	Chelsea
JF1	Anthony Taylor man of the match....? #FACupFinal #Arsenal #Chelsea	FACupFinal	Arsenal	Chelsea
Arlene	RT @tinkerpuss: The #FACupFinal is REAL sport. Greyhound racing is cr...	FACupFinal	arsenal	Chelsea
giuliogestione	RT @JimmyCase8: #WaTcH FA Cup Final live stream #Arsenal vs #Chel... HQD: https://t.co/rZhQnLMOx6 #FACupF...	WaTcH	Arsenal	Chelsea
Jack Gustard	For the majority of the game... #FACupFinal #Arsenal #Chelsea @mert...	FACupFinal	Arsenal	Chelsea

Ilustración 28: Datos filtrados

Fórmulas

Concluimos que la rivalidad es la siguiente:

$$R(C_{Arsenal}, C_{Chelsea}) = \text{ABS}\left(\frac{\sum_{i=1}^{n} S_i(\%\ de\ positividad)}{242} - 1\right) =$$

$$\text{ABS}\left(\frac{165{,}437}{242} - 1\right) = \mathbf{0{,}316376 = 31\%}$$

Utilizando todos los tweets referentes a esta comunidad, hemos almacenado unos 20.000 en total, se ha calculado la hostilidad que presenta la comunidad global que resulta de la agrupación de las anteriores comunidades, Arsenal, Chelsea, y el resto de seguidores que utilizan el hashtag FACupFInal.

$$H(FACupFinal) = \sum_{j=1}^{n} R\,(C_{Arsenal}, C_{Chelsea}, C_{Resto\ de\ comunidades},) = \mathbf{0{,}62 = 62\%}$$

Hashtag Bundesliga

Para esta comunidad se ha encontrado la dificultad de que pocos usuarios mencionaban a dos comunidades opuestas en sus tweets, por ello se ha decido analizar la rivalidad que presentan los seguidores del Bayern de Múnich por un lado y por otro los fans del equipo Borussia de Dortmund. De las 6192 ocurrencias que contiene la colección "Bundesliga", encontramos que 120 hacen alguna referencia al Borussia y 345 al Bayern lo que reduce bastante la cantidad de datos por comunidad.

Datos

user	texto	hashtag/0	hashtag/1
SocialCSS Bundesliga	https://t.co/HzvBkWWmBq updates background-image #71 https://t.c...	FCBayern	
RTS_Sport	.@philipplahm i @XabiAlonso ovacijama ispraćeni u igračku penziju &g...	FCBayern	Bundesliga
Just Soccer	Champions of the Bundesliga! We have a few Bayern jerseys left for o...	fcbayern	
Yari Edebaldo	En Alemania los hinchas del #BayernMunich festejan el titulo de la #Bu...	BayernMunich	Bundesliga
Rahul Rane	#FCBayern #bundesliga #Philipp Lahm #Xabi Alonso #champions	FCBayern	bundesliga
Football Pointer	Classy Philipp Lahm says auf Wiedersehen to Bayern Munich in style #...	bayernmunich	bundesliga
Football Pointer	Bayern Munich ease past Freiburg as Philipp Lahm and Xabi Alonso de...	bayernmunich	bundesliga
SocialCSS Bundesliga	https://t.co/HzvBkWWmBq updates card #68 https://t.co/kYgPXBzz0T...	FCBayern	Bundesliga
Bayern Munich Fans	#FCBayern #DieBayern #MiaSanMia Bayern Munich 4-1 Freiburg: Bun...	FCBayern	DieBayern
Asietta Freak✿	RT @SprockAnastacia: VIDEO» Vid & pics of the @AnastaciaMusi...	BayernMunich	Freiburg
Schalke_Canada FC	RT @BundesligaSpot: 2016/17 Bundesliga:	FCBayern	UCL

Ilustración 29: Datos filtrados Bayern

Fórmulas

Calculamos la rivalidad que presentan los seguidores del Bayern de Múnich:

$$\mathrm{R}\,(C_{Bayern}) = \mathrm{ABS}\left(\frac{\sum_{i=1}^{n} S_i(\%\ de\ positividad)}{345} - 1\right) = \mathbf{0{,}494322 = 49\%}$$

Calculamos la rivalidad de los seguidores del Borussia de Dortmund:

$$\mathrm{R}\,(\mathrm{C_{Borussia}}) = \mathrm{ABS}\left(\frac{\sum_{i=1}^{n} \mathrm{S_i}(\%\ \mathrm{de\ positividad})}{120} - 1\right) = \mathbf{0{,}7237 = 72\%}$$

Calculamos la hostilidad que presenta la comunidad total, que incluyen las comunidades anteriormente mencionadas y el resto de comunidades que aparezcan en los tweets, en este caso serían el resto de quipos participantes en la liga:

$$\mathrm{H}\left(C_{Bundesliga}\right) = \sum_{j=1}^{n} R\left(C_{Bayern}, C_{Borussia}, C_{Resto\ de\ comunidades},\right) =$$

$$\mathrm{ABS}\left(\frac{2128{,}309}{6192} - 1\right) = \mathbf{0{,}656281 = 65\%}$$

Comparamos la suma de la rivalidad de ambas comunidades frente a la hostilidad que presenta la comunidad global y comprobamos la relación de los resultados.

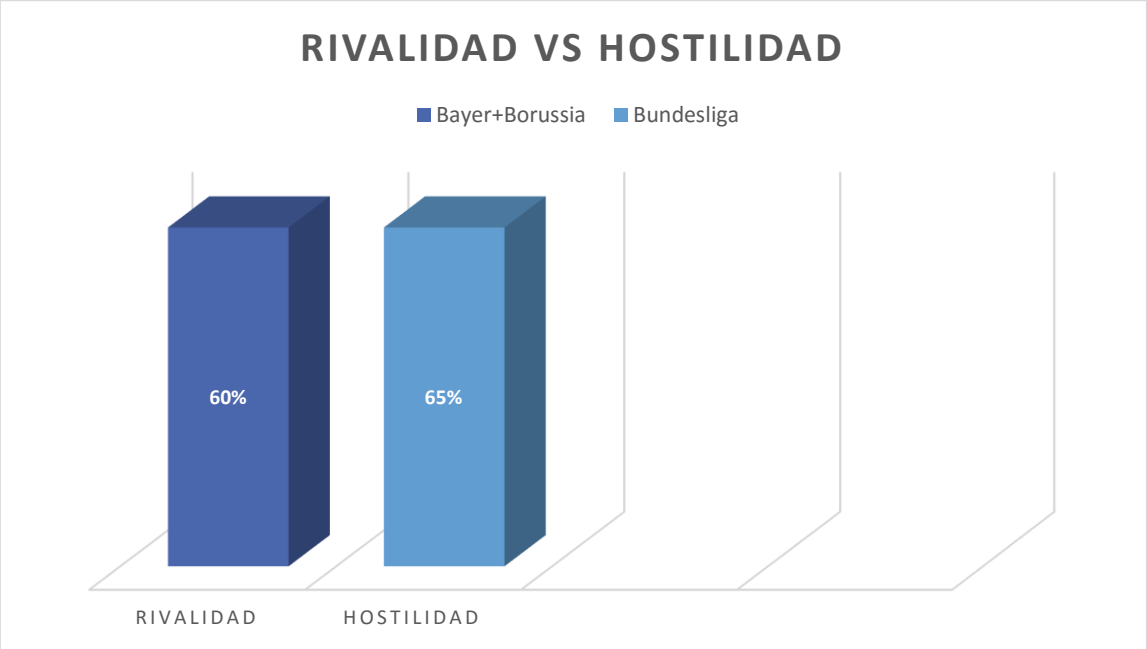


Ilustración 30: Comparativa rivalidad-hostilidad en liga Alemana

Hashtag Eurovisión

Datos

Separamos esta comunidad en países diferenciando entre España, Portugal y Reino Unido. Obtuvimos los siguientes datos filtrados:

- 200 tweets referidos a Reino Unido.
- 100 tweets referidos a España
- 400 tweets referidos a Portugal
- 60 tweets referidos a Irlanda
- 82 tweets referidos a Alemania

En la siguiente tabla expondremos la rivalidad entre las distintas comunidades:

	Reino Unido	**España**	**Portugal**	**Irlanda**	**Alemania**
Reino Unido	0	0,75	0,5	0,89	0,73
España	0,67	0	0,6	0,56	0,77
Portugal	0,45	0,64	0	0,5	0,6
Irlanda	0,89	0,67	0,56	0	0.76
Alemania	0,78	0,89	0,55	0,64	0

Tabla 21:Rivalidades Eurovisión

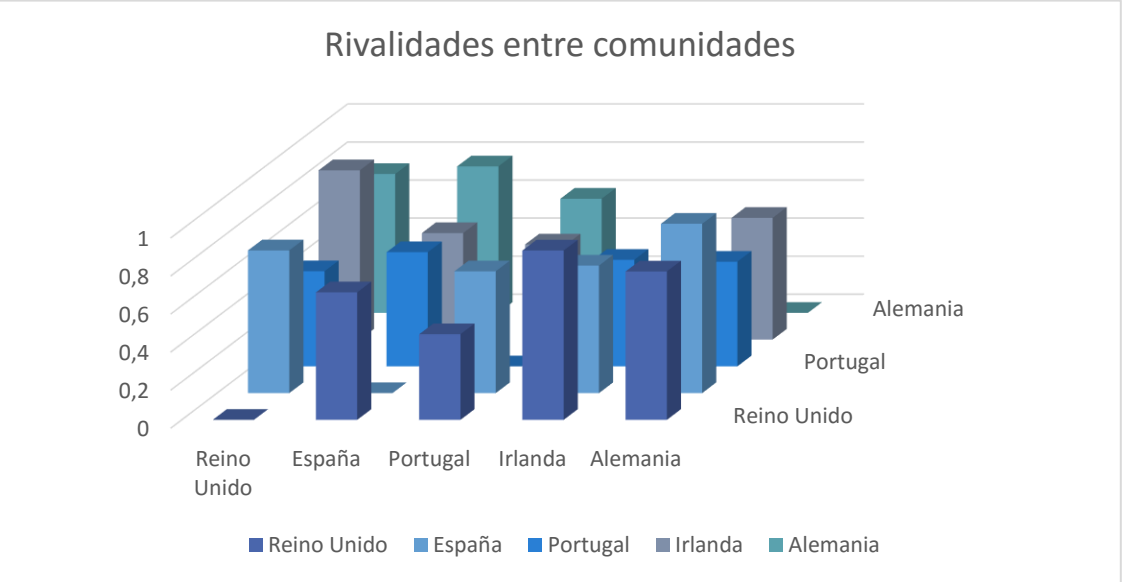


Ilustración 31: Rivalidades en Eurovision

Fórmulas

Calculamos la hostilidad de la comunidad global:

$$H\ (C_{Eurovision}) = \text{ABS}\left(\frac{1899{,}6}{3100} - 1\right) = \mathbf{0{,}387 = 38{,}7\ \%}$$

6.2. Estudio de los resultados obtenidos

Como podemos ver los valores para la liga alemana de fútbol son los siguientes:

Comunidad	**Rivalidad**
Bayern	49%
Borussia	72%
Hostilidad	65%

Si nos fijamos en el número de seguidores de ambas comunidades, podemos constatar que ambas son líderes en ellos dentro del conjunto de tweets que hemos analizado. Por lo tanto, un primer descubrimiento sería que la rivalidad tiene relación con las comunidades más populares.

Por otro lado, realizamos una comparación de tweets clasificados como positivos y aquellos como negativos, y vemos, como resulta lógico, que las comunidades que cuentan con más negatividad en sus tweets también son las más hostiles, sin embargo no queremos decir que la rivalidad sea en su totalidad una medición de la negatividad de una comunidad, puesto que también influye el grado de positividad que tengan los tweets. Es decir, que una comunidad "poco positiva" puede llegar a tener una rivalidad alta, sin ser necesario que la mayoría de sus tweets sean negativos.Pondremos un ejemplo para explicar este caso:

Con una pequeña muestra de unos 10 tweets , se compara el caso en el que seis de ellos son positivos con unos porcentajes en torno al 50 – 60 % , mientras que por otro lado se tiene otra muestra con 3 tweets positivos con porcentajes de positividad en torno al 80 – 90 % , los resultados de rivalidad son de 67% y 69 % respectivamente, lo que constata la conclusión anterior, puesto que son cantidades similares cuando se podría suponer que el segundo ejemplo debería tener una rivalidad muy alta por tener su mayoría de tweets negativos.

También hemos visto que la rivalidad tiene que ver con el vencedor de un evento, en este caso, la liga alemana fue ganada por el Bayern de Múnich, quedando en tercer puesto el Borussia de Dormunt,el cuál tiene una rivalidad mucho más alta.

Por otra parte, estudiando los resultados de la *tabla 7: Rivalidades Eurovisión,* concluiremos que países como Reino Unido e Irlanda están claramente enfrentados, la comunidad de los espectadores alemanes presenta una alta hostilidad en general hacia sus competidores, y el país ganador del concurso , Portugal, no presenta mucha rivalidad por parte de sus competidores, puesto que la gran mayoría de los tweets relacionados con Portugal son de enhorabuena por haber ganado, y por tanto son en su mayor parte de sentimiento positivo.

También se puede observar que rivalidad y hostilidad guardan una gran relación, puesto que las comunidades que presentan altos porcentajes de rivalidad entre ellas, contribuyen a que en el cálculo de la hostilidad este valor sea elevado también.

Comparando las hostilidades que presentan los grupos estudiados, se aprecia que los que presentan más antipatía son los relacionados con el ámbito deportivo y la política. El grupo menos hostil es el de Eurovisión. Con esto podemos concluir que las competiciones deportivas , en especial en el fútbol, generan bastante conflicto, mientras que los concursos musicales, como es Eurovision en este caso, parecen ser grupos más pacíficos.

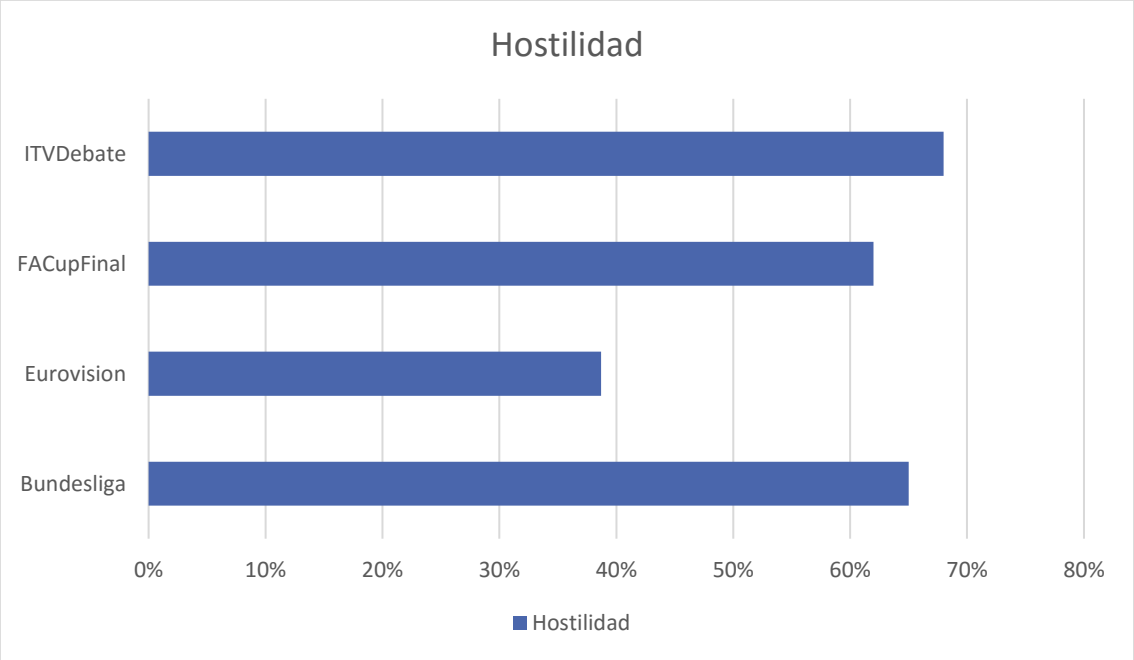


Ilustración 32: Hostilidad de las distintas comunidades

6.3. Pruebas unitarias

Realizaremos una serie de pruebas básicas con los datos y las herramientas que componen nuestro sistema para asegurar su correcto funcionamiento y calidad.

Código	PU-001
Descripción	Se comprueba que el sistema extrae tweets en tiempo real con un margen de ± 5 segundos de diferencia
Herramienta involucrada	Script y API twitter
Requisitos asociados	RF-001
Prueba	N/A

Tabla 22: PU-001

Código	PU-002
Descripción	Se verifica que la herramienta obtiene el número de tweets concreto que se le ha introducido como parámetro y además todos ellos contienen el hashtag que también ha sido introducido por parámetro
Herramienta involucrada	Script y API twitter
Requisitos asociados	RF-002,RF-003

Prueba	Se introducen los parámetros “MTV”, salida, y tres. Se crean tres tweets que contienen el hashtag “MTV”
	`>twitter_stream_download.py -q MTV -d salida -n 3`
	`{"created_at":"Sat Jun 03 16:06:39 +0000 2017","id":871035411333861376,"id_str":"871`
	`{"created_at":"Sat Jun 03 16:06:39 +0000 2017","id":871035411526701056,"id_str":"871`
	`{"created_at":"Sat Jun 03 16:06:40 +0000 2017","id":871035412701171712,"id_str":"871`

Tabla 23:PU-002

Código	PU-003
Descripción	Se comprueba que el fichero generado con los tweets está en formato JSON
Herramienta involucrada	Script y API twitter
Requisitos asociados	RF-004
Prueba	stream_MTV.json

Tabla 24: PU-003

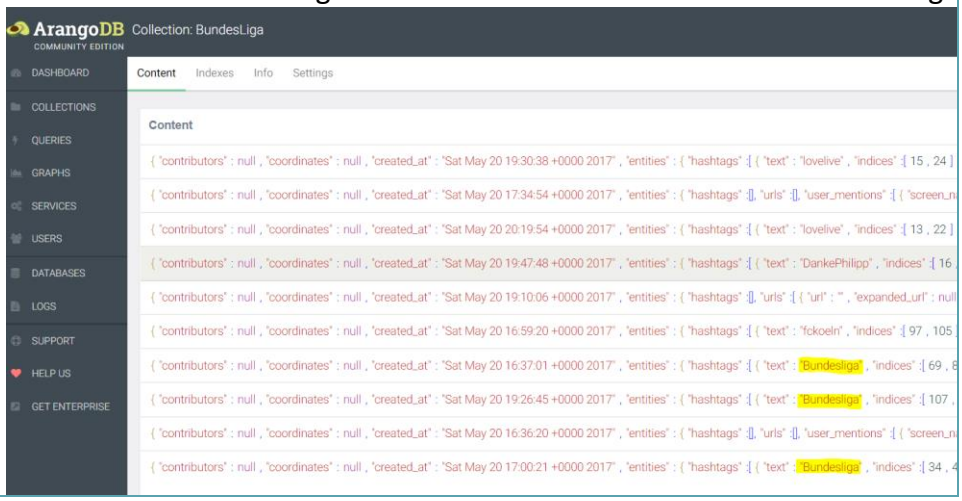
Código	PU-004
Descripción	Las colecciones de ArangoDB se componen de los documentos que hemos añadido mediante los comandos específicos.
Herramienta involucrada	ArangoDB
Requisitos asociados	N/A
Prueba	La colección "Bundesliga" contiene documentos con dicho hashtag 

Tabla 25:PU-004

Código	PU-005
Descripción	La recuperación de los datos requeridos dentro de un documento concreto coincide con los datos que muestra la herramienta
Herramienta involucrada	ArangoDB
Requisitos asociados	RF-005
Prueba	Se recuperan los tweets mostrando el usuario, texto, hashtag y menciones 

Tabla 26: PU-005

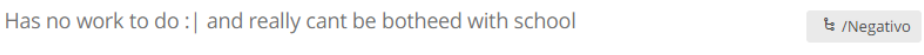
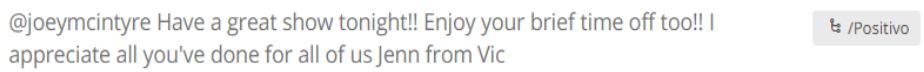
Código	PU-006
Descripción	Se comprueba que la clasificación del sentimiento de los tweets se corresponde con lo expuesto en el texto analizado
Herramienta involucrada	MonkeyLearn
Requisitos asociados	RF-006
Prueba	Tweet clasificado como negativo  Tweet positivo 

Tabla 27:PU-006

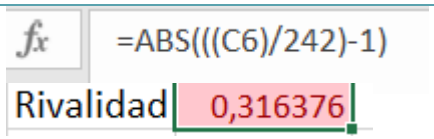
Código	PU-007
Descripción	Se comprueba que el cálculo de rivalidad produce un resultado en el rango esperado [0,1]
Herramienta involucrada	Excel
Requisitos asociados	RF-009
Prueba	

Tabla 28: PU-007

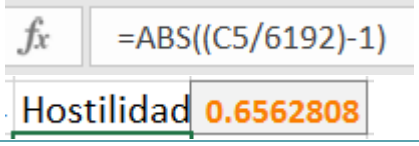
Código	PU-008
Descripción	Se comprueba que el cálculo de hostilidad produce un resultado en el rango esperado [0,1]
Herramienta involucrada	Excel
Requisitos asociados	RF-010
Prueba	

Tabla 29: PU-008

7. Planificación

La realización de un proyecto se puede dividir en tres grandes fases, planificación, ejecución y entrega o puesta en marcha. En este apartado definiremos la fase de planificación de nuestro proyecto, en la que se detalla por un lado una estimación inicial de las tareas que se preveía realizar y las horas que requerían, y por otro lado se compara con el tiempo que requirió realmente el proyecto.

Planificación inicial

Primero definiremos las tareas que deberíamos llevar a cabo para realizar el proyecto y las horas que estimamos que nos serían necesarias.

1. Propuesta y estudio de la idea inicial

Se realizaron varias reuniones con el tutor inicialmente para comentar y proponer ideas para la realización del proyecto. *Duración: 5 horas*

2. Análisis de la viabilidad del sistema

A continuación se estudiaron las ideas propuestas para determinar cuál de ellas era la más viable de cara a la realización del proyecto. *Duración: 10 horas*

3. Obtención de datos y procesamiento

Recolección de los datos necesarios para el proyecto. *Duración:25 horas*

4. Análisis del sistema

Se detallarán todas las partes que intervienen en el funcionamiento del sistema y se expondrán los requisitos que debe cumplir para su correcto funcionamiento. *Duración:24 horas*

5. Configuración del entorno tecnológico

Preparación de los programas y herramientas necesarias para el desarrollo del proyecto. *Duración: 35 horas*

6. Diseño e implantación del sistema

Duración:75 horas

7. Experimentación y pruebas

Se realizará un análisis de los datos que se han obtenido y se probará el sistema para asegurar su correcto funcionamiento y cumplimiento de los requisitos definidos. *Duración:45 horas*

8. Documentación

Realización de la memoria del proyecto y la presentación del mismo. *Duración:55 horas*

Ilustración 33: Horas estimadas

En total se han estimado unas 274 horas para la realización del proyecto, la Universidad Carlos III estima unas 300 horas por lo que nuestra estimación se aproxima a ello. En el gráfico se puede ver como se han repartido las horas en las diferentes tareas.

En el siguiente gráfico mostraremos las horas realmente invertidas en el proyecto, el total son 315 horas, algo más de lo que estimamos inicialmente.

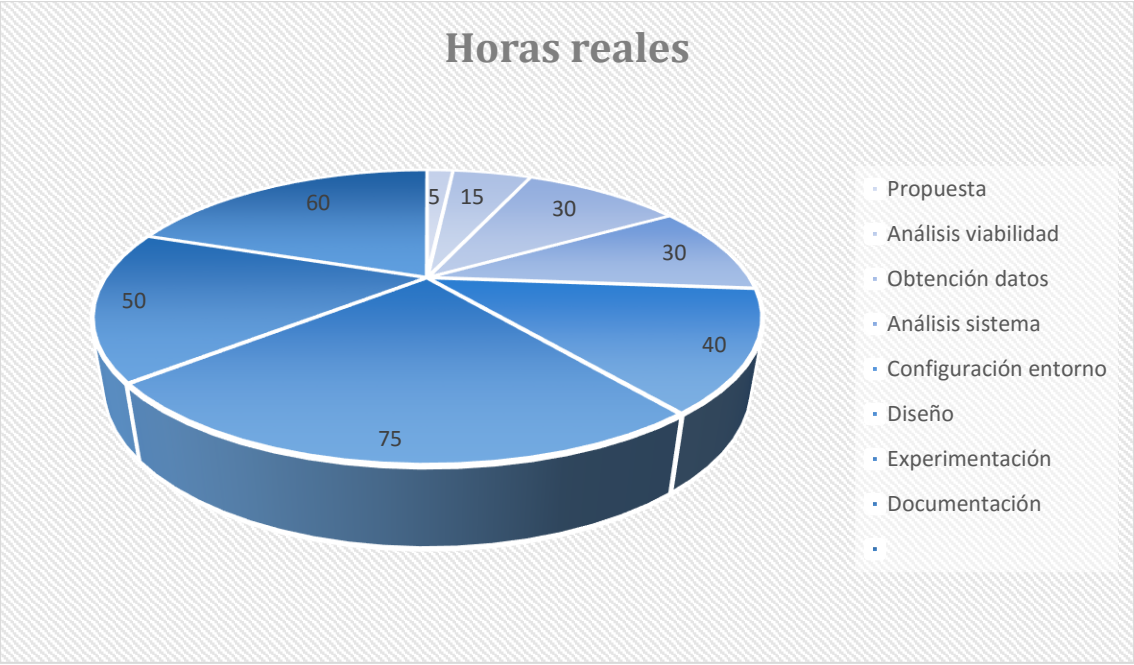


Ilustración 34: Horas reales

Por otra parte, hay que mencionar que la planificación de las tareas se hizo teniendo en cuenta que la realización del proyecto se debía compatibilizar con otras obligaciones como las prácticas en empresa y debido a los inconvenientes que puedan surgir se han hecho modificaciones a lo largo del desarrollo del mismo., tratando siempre de cumplir los objetivos propuestos.

Se muestra en un diagrama de Gantt la planificación seguida para dicho proyecto, el cual ha sido realizado con la herramienta OPENPROJ. En dicho diagrama aparece la duración del proyecto en días de trabajo, hemos determinado que cada día empleamos unas 3 horas de media en el proyecto para calcular cuántos días nos requerirá la planificación inicial desarrollada en horas.

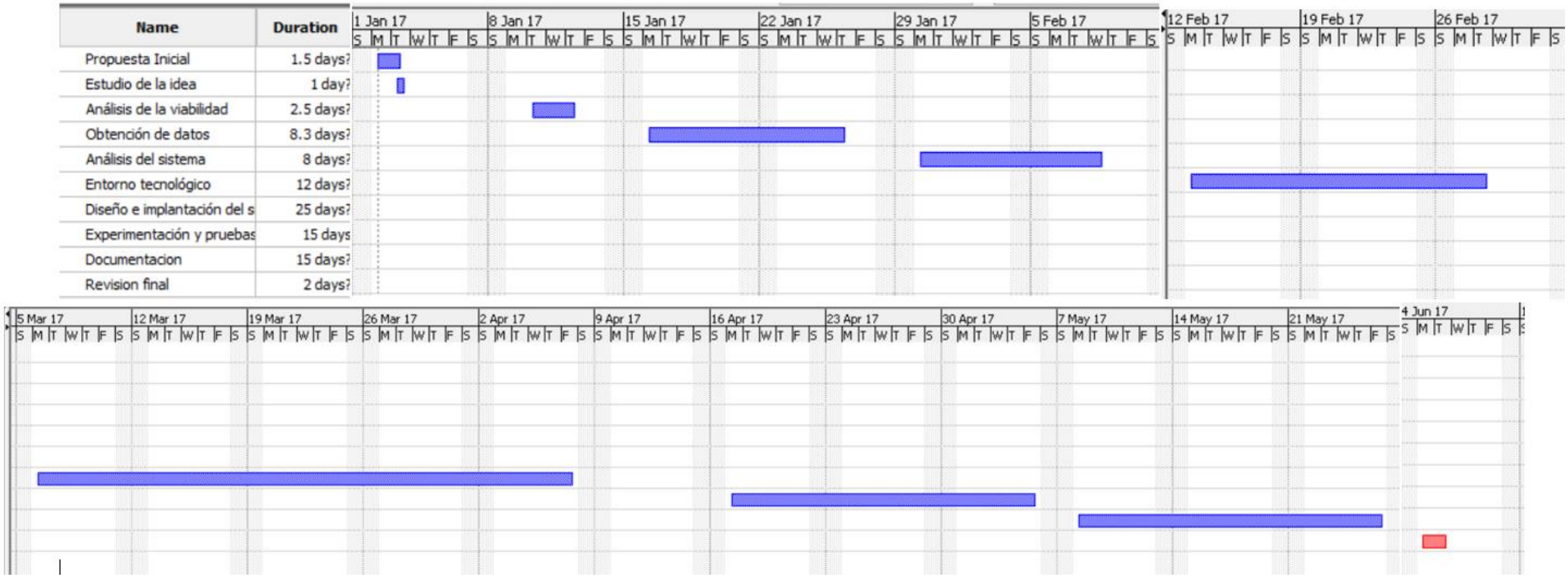


Ilustración 35:Diagrama de Gantt

8. Presupuesto

En este apartado se calculará el presupuesto necesario para la ejecución de nuestro proyecto, teniendo en cuenta tanto el coste de los materiales utilizados como el personal empleado. Este presupuesto será una estimación de lo que sería el coste real del proyecto y se supondrá que han intervenido en su realización un ingeniero senior, que será el tutor del proyecto, y un ingeniero junior que sería el estudiante que realiza el proyecto.

Realizando una estimación se determina que un ingeniero junior cobraría unos 20€/h mientras que uno senior llegaría a los 30€/h.

Concepto	Coste
Ingeniero Junior	300 horas x 20€/h = 6000€
Ingeniero senior	15 horas x 30€/h = 450€
TOTAL CAPITAL HUMANO	**6.450€**

Tabla 30:Cálculo capital humano

Por otra parte, se debe añadir el coste de las herramientas utilizadas para el proyecto, teniendo en cuenta el período de amortización que calcularemos de la siguiente manera [22]:

Para los equipos informáticos el porcentaje que corresponde es del 26%, por lo tanto, como hemos utilizado el ordenador durante unos 7 meses, su valor será 21,66 X 7 = 151,67 €.

Para los programas informáticos se aplica el mismo porcentaje, como todos los softwares son gratuitos menos Microsoft Office el cual hemos utilizado durante 7 meses, calcularemos su valor acorde a este tiempo.

Concepto	Precio	Amortización	Coste total
Portátil Toshiba	1000€	21,66 X 7	151,67€
Microsoft Office 2016	99,99€	2,16 X 7	15,12€
OpenProj	0€	-	0€
ArangoDB	0€	-	0€
Gephi	0€	-	0€
MonkeyLearn	0€	-	0€
TOTAL HERRAMIENTAS			**166,79€**

Tabla 31: Cálculo herramientas

Además, se debe tener en cuenta el IVA (21%) y el beneficio (20%), por lo que el presupuesto total del proyecto sería:

Concepto	Precio
Capital humano	6450€
Herramientas	166,79€
Beneficio 20%	1323,35 €
Total SIN IVA	**7940,14€**
Total CON IVA (21%)	**9607,57 €**

Tabla 32: Coste total del proyecto

9. Marco regulador

Para este proyecto se ha aplicado la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal [23], la cual tiene por objeto garantizar y proteger, en lo que concierne al tratamiento de los datos personales, las libertades públicas y los derechos fundamentales de las personas físicas, y especialmente de su honor e intimidad personal y familiar.

También hay que destacar que esta ley se ha visto modificada para ajustarse a los derechos de los usuarios en internet y podemos numerar una serie de cambios han sido impuestos por la Unión Europea [24]:

- Todas las personas que así lo deseen tienen derecho al olvido en Internet.
- Las empresas sólo podrán procesar datos personales con el consentimiento exclusivo de los usuarios.
- Se establece un nuevo derecho de "portabilidad de datos". Toda persona podrá solicitar a cualquier red social, que migre todos sus datos personales a otra red que considere más segura, como, por ejemplo, Twitter.
- Se prohíbe la creación de perfiles modificados (profiling). Es decir, no se podrán utilizar los datos personales cedidos por un usuario para otro fin que no sea el explícitamente acordado en un principio.

En relación a nuestro trabajo, se va a explicar brevemente cómo regula Twitter los datos personales con el fin de cumplir la anteriormente citada ley. Twitter cuenta con el consentimiento de sus usuarios para recopilar cualquier tipo de información que estos puedan proporcionar, podemos diferenciar tres tipos de datos que Twitter trata:

1. Datos que obtiene directamente del usuario: Desde el correo electrónico y una contraseña para darse de alta con un perfil, hasta datos de cesión voluntaria e imprescindible para el uso de aplicaciones dentro de Twitter. En este caso, las posibilidades de uso de la información dependerán de las políticas de privacidad y seguridad que el usuario consienta en su proceso de confirmación de alta.

2. Datos que se obtienen en abierto: El propio timeline de Twitter ofrece una serie de informes que la Agencia Española de Protección de Datos no establece como fuente de acceso público, ya que la LOPD no reconoce a Internet como un medio de comunicación propiamente dicho.

3. Datos que Twitter obtiene tras el permiso de usuario para conectarse a su cuenta: Su página de privacy ya lo indica así textualmente: "Twitter puede compartir o revelar la información del usuario con su consentimiento, como al utilizar la web de una tercera parte para acceder a la cuenta de Twitter".

"En todos los sentidos y, desde 2012, la Agencia Española de Protección de datos, reconoce que el mecanismo utilizado por Twitter, así como su política de privacidad, respetan lo acordado en el artículo 5.1 de la Ley Orgánica de Protección de Datos (LOPD)".

9.1. Impacto socio-económico

Como se explica en el informe de PWC [25], "los estudios de impacto económico sirven para medir la repercusión y los beneficios de inversiones en infraestructuras, organización de eventos, así como de cualquier otra actividad susceptible de generar un impacto socioeconómico, incluyendo cambios legislativos y regulatorios"

Es decir, se debe realizar un estudio para determinar los posibles beneficios y ventajas que generará nuestro proyecto, además de su posible repercusión en la sociedad actual. Realizar un análisis completo para determinar un impacto económico es una tarea compleja ya que de ello dependen multitud de variables, no sólo del tipo de activo que queremos valorar, si no de condiciones externas como el país o el momento temporal en el que nos encontramos.

Es por ello que realizaremos un análisis estimado de nuestro proyecto, exponiendo así el impacto que consideramos que creará tanto económica como socialmente.

Ilustración 36: Análisis del impacto [25]

En el caso del impacto económico, se plantearía el coste que tendría la realización de estudios y análisis de sentimiento en el caso de que dicho análisis tuviera que ser realizado manualmente por personas en lugar de utilizar el sistema desarrollado en nuestro trabajo. Para ello compararemos el cálculo que realizamos en la estimación del coste de nuestro proyecto frente a lo que costaría que un analista de datos empleara el tiempo que hemos requerido en realizar el proyecto, en realizar manualmente un análisis de sentimiento de los tweets y sus posteriores conclusiones, suponemos que tardaría el doble de lo que se ha tardado en realizar el proyecto.

Cálculo del sueldo de un Analista de datos → 40.000 €/ anuales / 14 pagas = 2857,14€

2857,14 / 20 días laborables X mes = 142,85 € x día 142,85/8= 17,85 €/h

Concepto	Precio
Sueldo Analista *17,85* €/h x 630 horas	11245,5
Posibles extras	500€
Beneficio	2349,1 €
Total Sin IVA	**14094,6€**
Total con IVA	**17054,46€**

Tabla 33:Coste del estudio realizado por humanos

Concepto	Precio
Precio proyecto	9607,57 €
Mantenimiento	500€
Total	**10.107,57 €**

Tabla 34: Coste del proyecto

La diferencia de costes entre ambos métodos es de 6946,89 €, es una cifra que representa un considerable ahorro entre un caso y otro, podemos concluir con que el impacto económico de nuestro proyecto sería positivo, puesto que el coste del desarrollo y mantenimiento de nuestra aplicación es menor que el coste que supondría realizar un análisis de datos por otras vías.

Por otra parte, nuestro trabajo también puede tener cierta repercusión social. Con nuestro sistema se podría llegar a determinar y conocer ciertos patrones de comportamiento en distintos grupos sociales. Puesto que el objetivo es captar las actitudes violentas u hostiles, se podría llegar a evitar o moderar las consecuencias negativas que se derivan de ellas, como por ejemplo casos de violencia física que se dan en ciertos eventos que hemos analizado en el trabajo como son los partidos de fútbol. De esta forma se podría clasificar los grupos socialmente violentos u hostiles y de ello se podrían realizar ciertas aplicaciones.

Llevando nuestro estudio al ámbito deportivo, por ejemplo, se podría prevenir el enfrentamiento que se da entre estos grupos sociales, aplicando medidas de seguridad y evitando el contacto entre los rivales.

Otra aplicación en el ámbito social sería la detección de casos de acoso o bullying, como hemos explicado antes, el acoso se ha ampliado ahora también al ámbito tecnológico, y mediante el análisis de la rivalidad entre usuarios cuyos comportamientos puedan resultar violentos u hostiles, podremos llegar a conocer si estos derivan en un caso de acoso a otra persona, pudiendo así tomar medidas correctivas hacia esas personas.

También se podría aplicar a casos de mejora en la ciberseguridad, puesto que en este ámbito se pretende detectar amenazas y proteger nuestros datos en la red, con nuestro sistema podríamos ligar ciertos comportamientos agresivos al posible de desarrollo de amenazas por parte de esos usuarios, diseñando así un sistema preventivo ante ataques en la red.

10. Conclusiones y trabajos futuros

10.1. Conclusiones

Para finalizar, concluiremos con que se han conseguido los objetivos que se plantearon inicialmente en el proyecto, conseguir medir la rivalidad y hostilidad en diferentes comunidades definidas en Twitter mediante un análisis de sentimiento.

Como partes críticas del proyecto, podemos destacar la obtención de los datos en Twitter, ya que, aunque aparentemente parezca una tarea sencilla por la gran cantidad de información de la que disponemos en la red, no fue tan fácil encontrar los datos concretos con los que podíamos trabajar, puesto que debían cumplir una serie de requisitos para que nuestro análisis tuviera éxito, como son el idioma, o que hubiera descriptores que diferenciaran comunidades dentro de Twitter.

También es un aspecto clave el análisis de sentimiento de los tweets, puesto que es de gran importancia a la hora de aplicar nuestras fórmulas de rivalidad y hostilidad y determinaran claramente el resultado de la misma. Aunque hay una gran cantidad de herramientas que clasifican textos por su sentimiento, se tomó la decisión de utilizar una que permitiera crear un modelo de predicción debido a que así podríamos conseguir datos más precisos y que se ajustaran más a nuestro objetivo de clasificación.

En cuanto a los resultados obtenidos, se han podido sacar conclusiones acerca de las comunidades rivales en Twitter como cuáles son los grupos sociales que presentan más conflictos, como es el caso del fútbol o la política. Además, hemos podido analizar los factores que intervienen a la hora de medir la antipatía entre grupos, como son su popularidad en Twitter, o que sean ganadores de cierta competición o concurso.

Además, durante el desarrollo del proyecto se han puesto en práctica conocimientos adquiridos durante la carrera, tales como el ciclo de vida del proyecto y sus diferentes fases desde la planificación hasta la ejecución y presentación del resultado final.

A nivel personal, me ha sorprendido todas las herramientas que hacen posible el análisis de datos en las redes sociales, así como todas las utilidades y aplicaciones que tiene, y sobre todo la gran cantidad de información que sin casi darnos cuenta exponemos diariamente en la red y de la cual se puede extraer muchas conclusiones, patrones de comportamiento, intereses por ciertos temas, nuestros gustos o por el contrario las cosas que nos desagradan.

10.2. Trabajos futuros

Se proponen una serie de mejoras como complemento al trabajo realizado:

- Mejora del modelo de predicción para una clasificación del sentimiento más precisa: Se podría desarrollar un modelo de predicción que contara con más ejemplos como entrenamiento, además de afinar los tweets que sirven como ejemplo clasificando estos con más detalle y eliminando palabras o caracteres vacíos que no influyan a la hora de determinar el sentimiento.
  En cuanto a las medidas de precisión y recall se trataría de acercar estos valores todo lo que fuera posible al 100% de manera que la exactitud del modelo también rondara este porcentaje, obteniendo así un modelo totalmente fiable y con pocos errores a la hora de clasificar la información.
  Por otra parte, se debe mencionar que el campo del análisis del sentimiento es amplio y complejo. Cada vez se desarrollan nuevas y mejores técnicas y algoritmos que permiten clasificar y detectar el sentimiento de los textos, sin embargo, hay que reconocer que es un campo muy subjetivo puesto que lo que para una persona puede reflejar un sentimiento, para otra puede ser todo lo contrario. Teniendo en cuenta esto, se propone también como mejora de la herramienta desarrollada, un clasificador del sentimiento mucho más preciso, que tuviera en cuenta más factores a la hora de detectar la polaridad, como por ejemplo, detección de ironía, sarcasmo, felicidad o tristeza, de manera que así conseguiríamos una clasificación mucho más concreta del sentimiento de los textos propuestos.
- Añadir más factores que determinen la relación entre dos comunidades: En nuestro estudio, detectamos que una comunidad tiene relación con otra mediante los hashtags utilizados, sin embargo, sería un gran avance poder determinar cómo están ligadas dos comunidades entre sí, analizando patrones que puedan llegar a determinar que ciertos usuarios se están refiriendo concretamente a otros grupos en sus tweets.
- Creación de un clasificador en español para poder realizar estudios sobre comunidades más diversas y no sólo de habla inglesa. La dificultad de encontrar contenido ya clasificado en otro idioma distinto del inglés, hace que nuestro trabajo se centre en comunidades de habla inglesa puesto que es la única forma de clasificar los tweets, como mejora se podría añadir un corpus de tweets en español y clasificarlos como positivos o negativos para poder añadirlos como ejemplos de entrenamiento y así conseguir un motor de clasificación en español.

11. Referencias

[1]. Enciclopedia de Clasificaciones (2017). "Tipos de redes sociales". Recuperado de:

http://www.tiposde.org/internet/87-tipos-de-redes-sociales/

[2]. HiperTextual “Historia de Twitter” por Miguel Ángel

https://hipertextual.com/archivo/2011/03/historia-twitter/

[3]. “¿Análisis de redes sociales, que es? “Por Julián Cárdenas

http://networksprovidehappiness.com/analisis-de-redes-sociales-es/

[4]. Facebook

https://es.wikipedia.org/wiki/Facebook

[5]. Historia de YouTube

http://www.cad.com.mx/historia_de_youtube.htm

[6]. “Twitter Analytics: La guía más completa en 2016” por David Soto

http://davidsotoro.com/twitter-analytics-la-guia-mas-completa/

[7]. “Cinco herramientas para analizar los sentimientos de los tweets” por JJ Velasco

https://hipertextual.com/archivo/2010/12/cinco-herramientas-para-analizar-los-sentimientos-de-los-tweets/

[8]. “La importancia de la analítica en redes sociales” por Gabriela Campos Torres

http://www.semanticwebbuilder.org.mx/es_mx/swb/La_importancia_de_la_analitica_en_redes_sociales

[9]. “Ciberacoso o ciberbullying” por Dr. Juan Moisés de la Serna, Doctor en Psicología

http://www.webconsultas.com/mente-y-emociones/trabajo-y-tiempo-libre/ciberacoso-o-ciberbullying-9723

[10]. Definición de rivalidad extraído de

http://www.definicionabc.com/general/rivalidad.php

[11]. Definición de hostilidad extraído de:

http://definicion.de/hostilidad/

[12]. “44 estadísticas de Twitter para 2016” por Kit Smith

https://www.brandwatch.com/es/2016/06/44-estadisticas-twitter-2016/

[13]. "Lo que siempre quiso saber del API de Twitter y nunca se atrevió a preguntar" por Mari Luz Congosto
http://www.barriblog.com/2010/07/lo-que-siempre-quiso-saber-del-api-de-twitter-y-nunca-se-atrevio-a-preguntar/

[14]. Twitter developer documentation
https://dev.twitter.com/overview/api/tweets

[15]. Definición de Bundesliga
https://es.wikipedia.org/wiki/Bundesliga_(Alemania)

[16]. "Opposition leaders clash over Brexit in first general election debate" by Adam Bienkov and Adam Payne
http://uk.businessinsider.com/itv-leaders-debate-blog-general-election-nicola-sturgeon-tim-farron-paul-nuttall-leanne-wood-caroline-lucas-2017-5

[17]. Python documentation
http://docs.python.org.ar/tutorial/3/real-index.html

[18]. ArangoDB documentation
https://docs.arangodb.com/3.1/Manual/index.html

[19]. Gephi
https://gephi.org/

[20]. MonkeyLearn Documentation
https://monkeylearn.com/docs/

[21]. Quora "How does multinomial Naive Bayes work"
https://www.quora.com/How-does-multinomial-Naive-Bayes-work

[22]. Cómo se calcula la amortización
http://www.serautonomo.net/%C2%BFcomo-se-calcula-la-amortizacion.html

[23]. Estudios de impacto económico por PWC
https://www.pwc.es/es/sector-publico/assets/brochure-estudios-impacto-economico.pdf

[24]. Noticias Jurídicas, Ley de Protección de Datos
http://noticias.juridicas.com/base_datos/Admin/lo15-1999.t1.html#t1

[25]. "Como se aplica la ley de protección de datos en Twitter"
http://toolows.es/blog/como-se-aplica-la-ley-de-proteccion-de-datos-en-twitter

12. Acrónimos

RDBMS -> relational database management system,

HTTP-> hypertext transfer protocol

API -> Application Programming Interface

ARS -> Análisis de redes sociales

CSV-> Comma Separated Values

JSON-> JavaScript Object Notation

SDK-> software development kit

GUI-> graphical user interface

AQL-> Arango Query Languague

PWC->Price Waterhouse Coopers

AFC-> Arsenal Football Club

CFC-> Chelsea Football Club